Machine Learning, 57, 145–175, 2004
© 2004 Kluwer Academic Publishers. Manufactured in The Netherlands.

Learning to Decode Cognitive States from Brain Images

TOM M. MITCHELL tom.mitchell@cmu.edu
REBECCA HUTCHINSON
RADU S. NICULESCU
FRANCISCO PEREIRA
XUERUI WANG
School of Computer Science, Carnegie Mellon University

MARCEL JUST
SHARLENE NEWMAN
Psychology Department, Carnegie Mellon University

Editors: Nada Lavrač, Hiroshi Motoda and Tom Fawcett

Abstract. Over the past decade, functional Magnetic Resonance Imaging (fMRI) has emerged as a powerful new instrument to collect vast quantities of data about activity in the human brain. A typical fMRI experiment can produce a three-dimensional image related to the human subject's brain activity every half second, at a spatial resolution of a few millimeters. As in other modern empirical sciences, this new instrumentation has led to a flood of new data, and a corresponding need for new data analysis methods. We describe recent research applying machine learning methods to the problem of classifying the cognitive state of a human subject based on fRMI data observed over a single time interval. In particular, we present case studies in which we have successfully trained classifiers to distinguish cognitive states such as (1) whether the human subject is looking at a picture or a sentence, (2) whether the subject is reading an ambiguous or non-ambiguous sentence, and (3) whether the word the subject is viewing is a word describing food, people, buildings, etc. This learning problem provides an interesting case study of classifier learning from extremely high dimensional (10^5 features), extremely sparse (tens of training examples), noisy data. This paper summarizes the results obtained in these three case studies, as well as lessons learned about how to successfully apply machine learning methods to train classifiers in such settings.

Keywords: scientific data analysis, functional Magnetic Resonance Imaging, high dimensional data, feature selection, Bayesian classifier, Support Vector Machine, nearest neighbor, brain image analysis

1. Introduction

The study of human brain function has received a tremendous boost in recent years from the advent of functional Magnetic Resonance Imaging (fMRI), a brain imaging method that dramatically improves our ability to observe correlates of neural brain activity in human subjects at high spatial resolution (several millimeters), across the entire brain. This fMRI technology offers the promise of revolutionary new approaches to studying human cognitive processes, provided we can develop appropriate data analysis methods to make sense of this huge volume of data. A twenty-minute fMRI session with a single human subject produces

a series of three dimensional brain images each containing approximately 15,000 voxels, collected once per second, yielding tens of millions of data observations.

Since its advent, fMRI has been used to conduct hundreds of studies that identify specific regions of the brain that are activated *on average* when a human performs a particular cognitive function (e.g., reading, mental imagery). The vast majority of this published work reports descriptive statistics of brain activity, calculated by averaging together fMRI data collected over multiple time intervals, in which the subject responds to repeated stimuli of some type (e.g., reading a variety of words).

In this paper we consider a different goal: training machine learning classifiers to automatically decode the subject's cognitive state at a single time instant or interval. The goal here is to make it possible to detect transient cognitive states, rather than characterize activity averaged over many episodes. This capability would clearly be useful in tracking the hidden cognitive states of a subject performing a single, specific task. Such classifier learning approaches are also potentially applicable to medical diagnosis problems which are often cast as classification problems, such as diagnosing Alzheimer's disease. While the approaches we discuss here are still in their infancy, and the more traditional approach of reporting descriptive statistics continues to dominate fMRI research, this alternative data analysis approach based on machine learning has already begun to gain acceptance within the neuroscience and medical informatics research communities (e.g., Strother et al., 2002; Cox & Savoy, 2003; Mitchell et al., 2003).

This problem domain is also quite interesting from the perspective of machine learning, because it provides a case study of classifier learning from extremely high dimensional, sparse, and noisy data. In our case studies we encounter problems where the examples are described by 100,000 features, and where we have less than a dozen, very noisy, training examples per class. Although conventional wisdom might suggest classifier learning would be impossible in such extreme settings, in fact we have found it is possible in this problem domain, by design of appropriate feature selection, feature abstraction and classifier training methods tuned to these problem characteristics.

In this paper we first provide a brief introduction to fMRI, then describe several fMRI data sets we have analyzed, the machine learning approaches we explored, and lessons learned about how best to apply machine learning approaches to the problem of classifying cognitive states based on single interval fMRI data.

2. Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a technique for obtaining three-dimensional images related to neural activity in the brain through time. More precisely, fMRI measures the ratio of oxygenated hemoglobin to deoxygenated hemoglobin in the blood with respect to a control baseline, at many individual locations within the brain. It is widely believed that blood oxygen level is influenced by local neural activity, and hence this blood oxygen level dependent (BOLD) response is generally taken as an indicator of neural activity.

An fMRI scanner measures the value of the fMRI signal (BOLD response) at all the points in a three dimensional grid, or image. In the studies described in this paper, a three

dimensional image is captured every 1, 1.5, or 0.5 seconds. We refer to the cells within this three-dimensional image as *voxels* (volume elements). The voxels in a typical fMRI study have a volume of a few tens of cubic millimeters, and a typical three dimensional brain image typically contains 10,000 to 15,000 voxels which contain cortical matter and are thus of interest. While the spatial resolution of fMRI is dramatically better than that provided by earlier brain imaging methods, each voxel nevertheless contains on the order of hundreds of thousands of neurons.

The temporal response of the fMRI BOLD signal is smeared over several seconds. Given an impulse of neural activity, such as the activity in visual cortex in response to a flash of light, the fMRI BOLD response associated with this impulse of neural activity endures for many seconds. It typically increases to a maximum after approximately four to five seconds, returning to baseline levels after another five to seven seconds. Despite this prolonged temporal response, some researchers (e.g., Menon et al., 1998) have reported that the relative timing of events can be resolved to within a few tens of milliseconds (e.g. to distinguish the relative timing of two flashes of light—one in the left eye and one in the right eye), providing hope that at least some temporal characteristics of brain function can be studied at subsecond resolution using fMRI.

A small portion of fMRI data is illustrated in figure 1. This figure shows data collected over a fifteen second interval during which the subject read a word, decided whether it was a noun or verb (in this case, it was a verb), then waited for another word. This data was sampled once per second for fifteen seconds, over sixteen planar slices, one of which is shown in the figure.

3. Related work analyzing fMRI data

Over recent years there has been a growing interest within the computer science community in data processing for fMRI. One popular style of processing involves using Generalized Linear Models (GLM) as in Friston et al. (1995a, 1995b) and Bly (2001). Here a regression is performed for each voxel, to predict the signal value at that voxel, based on properties of the stimulus. The degree to which voxel activity can be predicted from stimulus features is taken as an indication of the degree to which the voxel's activity is related to the stimulus. Notice this regression problem (predict voxel activity given the stimulus) is roughly the inverse of the problem we consider here (predict cognitive state given all voxel activities). Others have used t-statistics to determine relevant active voxels, and yet others have used more complex statistical methods to estimate parameters of the BOLD response in the presence of noise (Genovese, 1999).

Various methods for modelling time series have been used for analyzing fMRI data. For example, Hojen-Sorensen, Hansen, and Rasmussen (1999) used Hidden Markov Models (HMM) to learn a model of activity in the visual cortex resulting from a flashing light stimulus. Although the program was not told the stimulus, the on-off stimulus was recovered as the hidden state by the HMM.

A variety of unsupervised learning methods have also been used for exploratory analysis of fMRI data. For example, Goutte et al. (1998) discussed the use of clustering methods for fMRI data. One particular approach (Penny, 2001) involved the application of Expectation

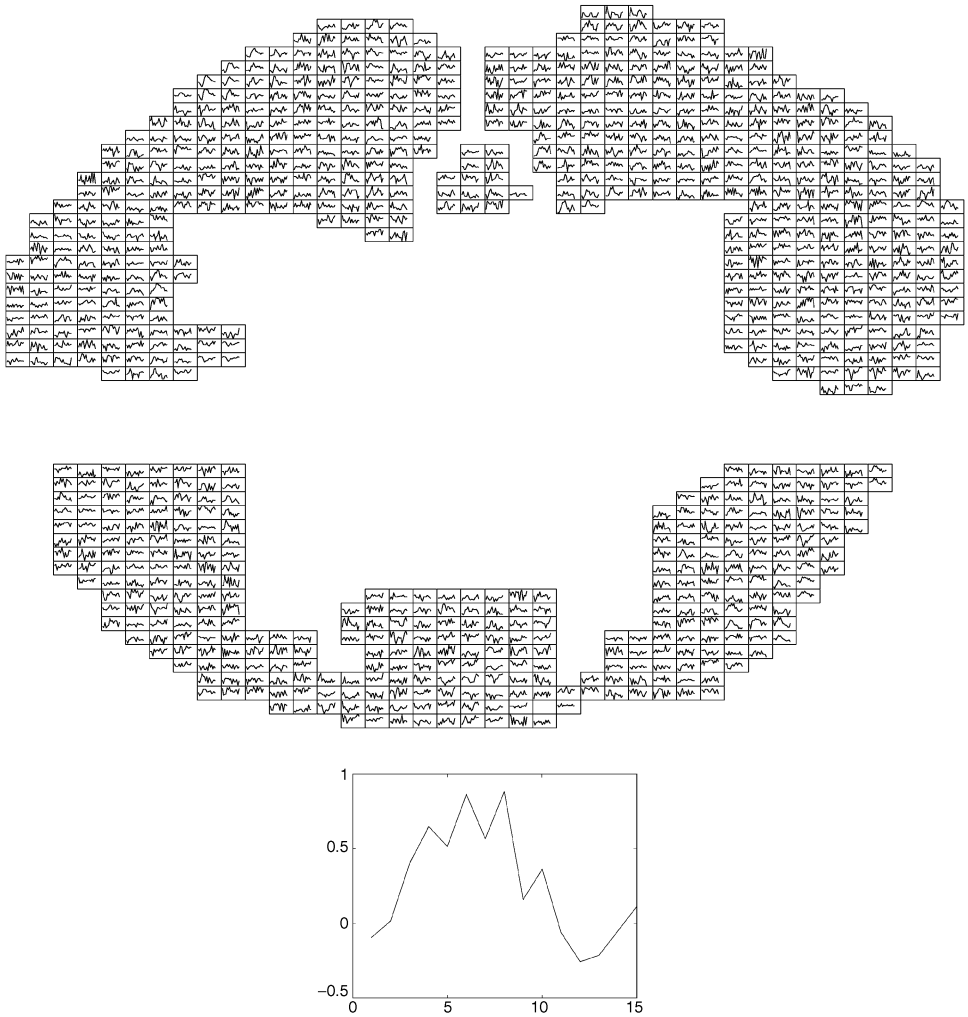

Figure 1. Typical fMRI data. The top portion of the figure shows fMRI data for a selected set of voxels in the cortex, from a two-dimensional image plane through the brain. A fifteen second interval of fMRI data is plotted at each voxel location. The anterior portion of the brain is at the top of the figure, posterior at bottom. The left side of the brain is shown on the right, according to standard radiological convention. The full three-dimensional brain image consists of sixteen such image planes. The bottom portion of the figure shows one of these plots in greater detail. During this interval the subject was presented a word, answered whether the word was a noun or verb, then waited for another word.

Maximization to estimate mixture models to cluster the data. Others have used Principle Components Analysis and Independent Components Analysis (McKeown et al., 1998) to determine spatial-temporal factors that can be linearly combined to reconstruct the fMRI signal.

While there has been little work on our specific problem of training classifiers to decode cognitive states, there are several papers describing work with closely related goals. For example, Haxby et al. (2001) showed that different patterns of fMRI activity are generated when a human subject views a photograph of a face versus a house, versus a shoe, versus a chair. While they did not specifically use these discovered patterns to classify subsequent single-event data, they did report that by dividing the fMRI data for each photograph category into two samples, they could automatically match the sample means related to the same category. More recently (Cox & Savoy, 2003) applied Support Vector Machine and Linear Discriminant Analysis to a similar set of data to successfully classify patterns of fMRI activation evoked by the presentation of photographs of various categories of objects. Others (Wagner et al., 1998) reported that they have been able to make better-than-random predictions regarding whether a visually presented word will be remembered later, based on the magnitude of activity within certain parts of left prefrontal and temporal cortices during that presentation.

In addition to work on fMRI, there has been related recent work applying machine learning methods to data from other devices measuring brain activity. For example, Blankertz, Curio, and Mller (2002) describe experiments training classifiers of brain states for single trial EEG data, while (Kjems et al., 2002; Strother et al., 2002) report training brain state classifiers for images obtained via Positron Emission Tomography (PET).

The work reported in the current paper builds on our earlier research described in Mitchell et al. (2003) and Wang, Hutchinson, and Mitchell (2003).

4. Approach

This section briefly describes our approach to data preprocessing, training classifiers, and evaluating them.

4.1. Data acquisition and preprocessing

In the fMRI studies considered here, data were collected from normal students from the university community. Typical studies involved between five and fifteen subjects, and we generally selected a subset of these subjects with the strongest, least noisy fMRI signal to train our classifiers. Data were preprocessed to remove artifacts due to head motion, signal drift, and other sources, using the FIASCO program (Eddy et al., 1998).[1] All voxel activity values were represented by the percent difference from their mean value during rest conditions (when the subject is asked to relax, and not perform any particular task). These preprocessed images were used as input to our classifiers.

In several cases, we found it useful to identify specific anatomically defined regions of interest (ROIs) within the brain of each subject. To achieve this, two types of brain images were collected for each subject. The first type of image, which has been discussed up to this point in the paper, captures brain activation via the BOLD response, and is referred to as a *functional image*. The second type of image, called a *structural image*, captures the static physical brain structure at higher resolution. For each subject, this structural image was used to identify the anatomical regions of interest, using the parcellation scheme of

Caviness et al. (1996) and Rademacher et al. (1992). For each subject, the mean of their functional images was then co-registered to the structural image, so that individual voxels in the functional images could be associated with the ROIs identified in the structural image.

4.2. *Learning methods*

In this paper we explore the use of machine learning methods to approximate classification functions of the following form

$$f : \text{fMRI-sequence}(t_1, t_2) \rightarrow \text{CognitiveState}$$

where fMRI-sequence(t_1, t_2) is the sequence of fMRI images collected during the contiguous time interval $[t_1, t_2]$, and where CognitiveState is the set of cognitive states to be discriminated. Each of our data sets includes fMRI data from multiple human subjects. Except where otherwise noted, we trained a separate classification function for each subject.

We explored a variety of methods for encoding fMRI-sequence(t_1, t_2) as input to the classifier. In some cases we encoded the input as a vector of individual voxel activities, a different activity for each voxel and for each image captured during the interval $[t_1, t_2]$. This can be an extremely high dimensional feature vector, consisting of hundreds of thousands of features given that a typical image contains 10,000 to 15,000 voxels, and a training example can include dozens of images. Therefore, we explored a variety of approaches to reducing the dimension of this feature vector, including methods for feature selection, as well as methods that replace multiple feature values by their mean. These feature selection and feature abstraction methods are described in detail in Section 6.3.

We explored a number of classifier training methods, including:

– *Gaussian Naive Bayes* (*GNB*). The GNB classifier uses the training data to estimate the probability distribution over fMRI observations, conditioned on the subject's cognitive state. It then classifies a new example $\vec{\mathbf{x}} = \langle x_1 \ldots x_n \rangle$ by estimating the probability $P(c_i \mid \vec{\mathbf{x}})$ of cognitive state c_i given fMRI observation $\vec{\mathbf{x}}$. It estimates this $P(c_i \mid \vec{\mathbf{x}})$ using Bayes rule, along with the assumption that the features x_j are conditionally independent given the class:

$$\hat{P}(c_i \mid \vec{\mathbf{x}}) = \frac{\hat{P}(c_i) \prod_j \hat{P}(x_j \mid c_i)}{\sum_k \left[\hat{P}(c_k) \prod_j \hat{P}(x_j \mid c_k) \right]}$$

where $\hat{P}$ denotes distributions estimated by the GNB from the training data. Each distribution of the form $\hat{P}(x_j \mid c_i)$ is modelled as a univariate Gaussian, using maximum likelihood estimates of the mean and variance derived from the training data. Distributions of the form $\hat{P}(c_i)$ are modelled as Bernoulli, again using maximum likelihood estimates based on training data. Given a new example to be classified, the GNB outputs posterior probabilities for each cognitive state, calculated using the above formula.

We considered two variants of the GNB, which differ only in their approach to estimating the variances of the univariate Gaussian distributions $\hat{P}(x_j \mid c_i)$. In

GNB-SharedVariance, it is assumed that the variance of voxel x_j is identical for all classes c_i. This single variance is estimated by the sample variance of the pooled data for x_j taken from all classes (with the class mean subtracted out of each value). In GNB-DistinctVariance, the variance is estimated separately for each voxel and class.

– *Support Vector Machine* (*SVM*). We used a linear kernel Support Vector Machine (see, for instance, Burges (1998)).
– *k Nearest Neighbor* (*kNN*). We use k Nearest Neighbor with a Euclidean distance metric, considering values of 1, 3, 5, 7 and 9 for k (see, for instance, Mitchell (1997)).

4.3. Evaluating results

Trained classifiers were evaluated by their cross-validated classification error when learning boolean-valued classification functions. When more than two classes are involved, our classifiers output a rank-ordered list of the potential classes from most to least likely. In this case, we scored the success of each prediction by the normalized rank of the correct class in this sorted list, which we refer to as the *normalized rank error*. Thus, the normalized rank error ranges from 0 when the correct class is ranked most likely, to 1 when it is ranked least likely. Note this normalized rank error is a natural extension of classification error when multiple classes are involved, and is identical to classification error when exactly two classes are involved. Note also that random guessing yields an expected normalized rank error of 0.5 regardless of the number of classes under consideration.

To evaluate classifiers, we generally employ k-fold cross-validation, leaving out one example per class on each fold. In the data sets considered in this paper, the competing classes are balanced (i.e., the number of available examples is the same for each competing class). Thus, by leaving out one example per class we retain a balanced training set for each fold, which correctly reflects the class priors.

Because the fMRI BOLD response is blurred out over several seconds, a strict leave-out-one-example-per-class evaluation can sometimes produce optimistic estimates of the true classifier error. The reason is straightforward: when holding out a test image occurring at time t, the training images at times $t + 1$ and $t - 1$ will be highly correlated with this test image. Therefore, if the images at $t - 1$ and $t + 1$ belong to the same class as the image at t, and are included in the training set, this can lead to optimistically biased error estimates for the held out example. When faced with this situation (i.e., in the Semantic Categories study described below), we avoid the optimistic bias by removing from the training set all images that occur within 5 seconds of the held out test image. In this case, our cross validation procedure involves holding out one test example per class, and also removing temporally proximate images from the training set.

5. Case studies

This section describes three distinct fMRI studies, the data collected in each, and the classifiers trained for each. In this section we summarize the success of the best classifier obtained for each of these studies. Section 6 discusses more generally the lessons learned across these three case studies.

5.1. Picture versus sentence study

In this fMRI study (Keller, Just, & Stenger, 2001), subjects experienced a collection of trials. During each trial they were shown in sequence a sentence and a simple picture, then answered whether the sentence correctly described the picture. We used this data to explore the feasibility of training classifiers to distinguish whether the subject is examining a sentence or a picture during a particular time interval.

In half of the trials the picture was presented first, followed by the sentence. In the remaining trials, the sentence was presented first, followed by the picture. In either case, the first stimulus (sentence or picture) was presented for 4 seconds, followed by a blank screen for 4 seconds. The second stimulus was then presented for up to 4 seconds, ending when the subject pressed the mouse button to indicate whether the sentence correctly described the picture. Finally, a rest period of 15 seconds was inserted before the next trial began. Thus, each trial lasted approximately 27 seconds. Pictures were geometric arrangements of the symbols $+$, $*$ and/or \$, such as

$$\begin{array}{c} + \\ \hline * \end{array}$$

Sentences were descriptions such as "It is true that the plus is below the dollar." Half of the sentences were negated (e.g., "It is not true that the star is above the plus.") and the other half were affirmative sentences.

Each subject was presented a total of 40 trials as described above, interspersed with ten additional rest periods. During each of these rest periods, the subject was asked to relax while staring at a fixed point on the screen. fMRI images were collected every 500 msec.

The learning task we consider for this study is to train a classifier to determine, given a particular 8-second interval of fMRI data, whether the subject is viewing a sentence or a picture during this interval. In other words, we wish to learn a separate classifier for each subject, of the following form

$$f : \text{fMRI-sequence}(t_0, t_0 + 8) \rightarrow \{\text{Picture, Sentence}\}$$

where t_0 is the time of stimulus (picture or sentence) onset. Thus, the input to the classifier is an 8-second interval of fMRI data beginning when the picture or sentence is first presented to the subject. Although the stimulus was presented for a maximum duration of only 4 seconds, we chose this 8-second interval in order to capture the full fMRI activity associated with the stimulus (recall from Section 2 that the fMRI BOLD signal often extends for 9–12 seconds beyond the neural activity of interest).[2]

There were a total of 80 examples available from each subject (40 examples per class). The fMRI-sequence was itself described by the activities of all voxels in cortex. The average number of cortex voxels per subject was approximately 10,000, and varied significantly by subject, based in large part on the size of the subject's head. Note that the eight second interval considered by the classifier contains 16 images (images were captured twice per second), yielding an input feature vector containing approximately 160,000 features, before feature selection.

The expected classification error of the default classifier (guessing the most common class) is 0.50 in this case. The average error obtained for the most successful trained classifier, using the most successful feature selection strategy, was 0.11, averaged over 13 subjects, with the best subject reaching 0.04 (refer to Section 6.2 for more details). These results are statistically highly significant, and indicate that it is indeed possible to train classifiers to distinguish these two cognitive states reliably.

In addition to these single-subject classifiers, we also experimented with training classifiers that operate across multiple subjects. In this case, we evaluated the classification error using a leave-one-subject-out regime in which we held out each of the 13 subjects in turn while training on the other 12. The mean error over the held out subject for the most successful combination of feature selection and classifier was 0.25. Again, this is significantly better than the expected 0.5 error from the default classifier, indicating that it is possible to train classifiers for this task that operate on human subjects who where not part of the training set. These results are described in detail in Section 6.4.

5.2. *Syntactic ambiguity study*

In this fMRI study (see Mason et al., 2004) subjects were presented with sentences, some of which were ambiguous, and were asked to respond to a yes-no question about the content of each sentence. The questions were designed to ensure that the subject was in fact processing the sentence. The learning task for this study was to distinguish whether the subject was currently reading an ambiguous sentence (e.g., "The experienced soldiers warned about the dangers conducted the midnight raid.") or an unambiguous sentence (e.g., "The experienced soldiers spoke about the dangers before the midnight raid.").[3]

Ten sentences of each of type were presented to each subject. Each sentence was presented for 10 seconds. Next a question was presented, and the subject was given 4 seconds to answer. After the subject answered the question, or 4 seconds elapsed, an "X" appeared on the screen for a 12 second rest period. The scanner collected one fMRI image every 1.5 seconds.

We are interested here in learning a classifier that takes as input an interval of fMRI activity, and determines which of the two types of sentence the subject is reading. Using our earlier notation, for each subject we trained classifiers of the form

$$f : \text{fMRI-sequence}(t_0 + 4.5, t_0 + 15) \rightarrow \text{SentenceType}$$

where SentenceType $= \{\text{Ambiguous, Unambiguous}\}$, and where t_0 is the time at which the sentence is first presented to the subject. Note the classifier input describes fMRI activity during the interval from 4.5 to 15 seconds following initial presentation of the sentence. This is the interval during which the fMRI activity is most intense.

There were a total of 20 examples for each subject (10 examples per class). The fMRI-sequence was described using only the voxels from four ROIs considered to be most relevant by a domain expert. These 4 ROIs contained a total of 1500 to 3508 voxels, depending on the subject. Note the 10.5 second interval considered by the classifier contains 8 images (images were captured every 1.5 seconds), yielding a classifier input vector containing from 12,000 to 28,064 features, depending on the human subject, before feature selection.

The expected classification error of the default classifier (guessing the most common class) in this case is 0.50, given the equal number of examples from both classes. The average error obtained by the most successful combination of feature selection and classifier was 0.25, averaged over 5 subjects, with the best single-subject classifier reaching an error of 0.10 (refer to Section 6.2 for more details).

5.3. Semantic categories study

In this study, 10 subjects were presented with individual nouns belonging to twelve distinct semantic categories (e.g., Fruits, Tools), and asked to determine whether the word belonged to a particular category. We used this data to explore the feasibility of training classifiers to detect which of the semantic categories of word the subject was examining.

The trials in this study were divided into twelve blocks. In each block, the name of a semantic category was first displayed for 2 seconds. Following this, the subject was shown a succession of 20 words, each presented for 400 msec and followed by 1200 msec of blank screen. After each word was presented, the subject clicked a mouse button to indicate whether the word belonged to the semantic category named at the beginning of the block. In fact, nearly all words belonged to the named category (half the blocks contained no out-of-category words, and the remaining blocks contained just one out-of-category word). A multi-second blank screen rest period was inserted between each of the twelve blocks. The twelve semantic categories of words presented were "fish," "four-legged animals," "trees," "flowers," "fruits," "vegetables," "family members," "occupations," "tools," "kitchen items," "dwellings," and "building parts." Words were chosen from lists of high frequency words of each category, as given in Battig and Montague (1968), in order to avoid obscure or multiple-meaning words. fMRI images were acquired once per second.

The learning task we considered for this study is to distinguish which of the twelve semantic categories the subject is considering, based on a single observed fMRI image. Following our earlier notation, we wish to learn a classifier of the form:

$$f : \text{fMRI}(t) \rightarrow \text{WordCategory}$$

where fMRI(t) is a single fMRI image, and where WordCategory is the set of 12 semantic categories described above.

A total of 384 example images were collected for each subject (32 examples per class, times 12 classes). All voxels from 30 ROIs were used, yielding a total of 8,470 to 11,136 voxels, depending on the subject. In this case the classifier input is a single image, so the classifier input dimension is equal to the number of voxels, prior to feature selection.

The trained classifier outputs a rank-ordered list of the 12 categories, ranked from most to least probable. We therefore evaluate classifier error using the normalized rank error described in Section 4, where the default classifier (guessing the most frequent class) yields an expected normalized rank error of 0.50. The normalized rank error for the most successful combination of feature selection and classifier was 0.08 (i.e. on average the correct word category was ranked first or second out of the twelve categories), over 10 subjects, with the best subject reaching 0.04. (refer to Section 6.2 for details).

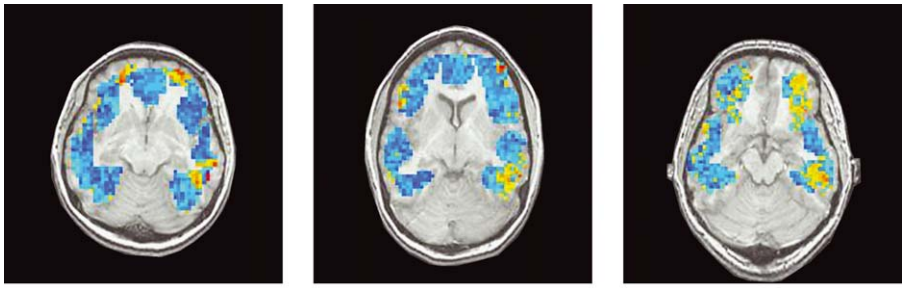

Figure 2. Color plots show locations of voxels that best predict the word semantic category, for three different human subjects. For each voxel, the color indicates the normalized rank error over the test set, for a GNB classifier based on this single voxel. Note the spatial clustering of highly predictive voxels, and the similar regions of predictability across these three subjects. The range of normalized rank errors is [Red $\approx$ 0.25, Dark Blue $\approx$ 0.6], with other colors intermediate between these two extremes. Each image corresponds to a single two-dimensional plane through the brain of one subject.

One reasonable question that can be raised regarding these classifier results is whether the classifier is indeed learning the pattern of brain activity predictive of semantic categories, or whether it is instead learning patterns related to some other time-varying phenomenon that influences fMRI activation. One unfortunate property of the experimental protocol for collecting data, from this point of view, is that all of the words belonging to a single category are presented within a single time interval (i.e., a single experiment block). In fact we do believe this temporal adjacency may be influencing our results, but we also believe the classifier is indeed capturing regularities primarily related to semantic categories. One strong piece of supporting evidence is that classifiers trained for different human subjects tend to rely on the same brain locations to make their predictions, and that these regions have been reported by others as related to semantic categorization. Figure 2 illustrates the brain regions containing the most informative fMRI signal for classification, across three subjects. In this figure, red and yellow indicate the voxels whose activity allows most accurate classification. Note the highly discriminating voxels are clustered together, in similar regions across these subjects. These locations for discriminability match those reported in earlier work on semantic categorization by Chao, Haxby, and Martin (1999), Chao, Weisberg, and Martin (2002), Ishai et al. (1999) and Aguirre, Zarahn, and D'Esposito (1998), as well as some novel areas that are currently under investigation.

6. Lessons learned

6.1. Can one learn to decode mental states from fMRI?

The primary goal leading to this research was to determine whether it is feasible to use machine learning methods to decode mental states from single interval fMRI data. The successful results reported above for all three data sets indicate that this is indeed feasible in

a variety of interesting cases. However, it is important to note that while our empirical results demonstrate the ability to successfully distinguish among a predefined set of states occurring at specific times while the subject performs specific tasks, they do not yet demonstrate that trained classifiers can reliably detect cognitive states occurring at arbitrary times while the subject performs arbitrary tasks. While our current results may already be of use in cognitive science research, we intend to pursue this more general goal in future work.

We also attempted but failed to train successful classifiers for several other classification functions defined over these same data sets. For example, we were unable to train an accurate classifier to distinguish the processing of negated versus affirmative sentences in the Picture versus Sentence study. We were also unsuccessful in attempts to train classifiers to distinguish the processing of true versus false sentences, or sentences which the subject answered correctly versus incorrectly. It may be that these failures could be reversed given larger training sets or more effective learning algorithms. Alternatively, it may be the case that the fMRI data simply lacks the information needed to make these distinctions. This line of research is still very new, and while the above results demonstrate the feasibility of discriminating a variety of cognitive states based on fMRI, at this point the question of exactly which cognitive states can be reliably discriminated remains an open empirical question. Given our initial successes plus those reported in Cox and Savoy (2003), as well as likely advances in brain imaging technology and likely progress in developing machine learning methods specifically for this type of application, we are optimistic that over time we will be able to decode an increasingly useful collection of cognitive states in an increasingly open ended set of experimental settings.

Interestingly, we found that the accuracy of our single-subject classifiers varied significantly among subjects, even within the same study. For example, when training an SVM for the 13 different subjects in the "Picture versus Sentence" study, the error rates of the 13 single-subject classifiers were 0.01, 0.04, 0.06, 0.06, 0.06, 0.06, 0.08, 0.10, 0.16, 0.18, 0.19, 0.20, and 0.25 (additional single-subject data is reported in Tables 2–4). Subjects producing high accuracies with one classifier were typically the same who produced high accuracies for other classifiers, and it is likely that the data for the worst-performing classifiers were corrupted by various kinds of noise (e.g., significant head motion during imaging). Considerable subject-to-subject variation in fMRI responses has been widely reported in the fMRI literature.

Before leaving the topic of whether one can train classifiers in this domain, it is worth considering the question of whether our reported better-than-random classifier error rates are simply the result of having tried many learning algorithms and feature selection methods, then reporting results for the approach that worked best. This is a general issue whenever one experiments with many learning approaches, then reports accuracy for the one that performs best over the test data. In the following subsections we describe in detail the different feature selection methods and learning methods we explored. For the purposes of this discussion, however, there are two important points to be made. First, we found that *every* learning algorithm (GNB, SVM, and *k*NN) produced significantly better than random classification accuracies for every case study, supporting our conclusion that it is feasible to learn classifiers in this domain. Second, when performing feature selection, features were chosen over a training set distinct from the test data, using a leave-one-out

cross validation approach, so that each test example had no influence on which features were selected. The only way that the test data influenced feature selection was in choosing the number of features, n, for which results are reported. This single parameter n was chosen for each learning algorithm and study, to maximize the *mean* accuracy over all single-subject classifiers trained by that algorithm for that study. Given that the number of single-subject classifiers ranged from five to thirteen, depending on the study, we conjectured that the choice of this single parameter n would exert only a very minor influence and that we could safely consider our reported accuracies to be very close to true accuracies.

To test this conjecture we conducted an experiment to compare this biased estimate (using the value of n which maximized test set performance) to an unbiased estimate obtained using a more elaborate and computationally intensive approach. To obtain the fully unbiased estimate, we employed a nested cross validation approach to partition the data repeatedly into three sets: a training set, optimization set, and final test set. The training and optimization sets were used to train a GNB classifier, to choose the feature selection method, and to choose the number n of features to be included. The final test set had no influence on training or selecting what to report, and was used only to provide a final unbiased estimate of accuracy. We ran this experiment over the data from all 13 subjects in our "Sentence versus Picture" data set. The resulting unbiased error estimate provided by the final test set was 0.183, whereas the original biased estimate provided by our standard two-set approach was 0.182. Furthermore, the number of features n selected by these two approaches were nearly identical. This experimental outcome supports our conjecture that our reported classifier accuracies, while in theory slightly biased, are in fact very close to their true accuracies.

6.2. *Which classifier works best?*

As discussed earlier, we experimented with three classifier learning methods: a Gaussian Naive Bayes (GNB) classifier, k-nearest neighbor (kNN), and linear Support Vector Machines (SVM). These classifiers were selected because they have been used successfully in other applications involving high dimensional data. For example, Naive Bayes classifiers, kNN, and SVM have all been used for text classification problems (Nigam et al., 2000; Joachims, 2001; Yang, 1999), where the dimension of the data is approximately 10^5, corresponding to the size of the natural language vocabulary.

To test the relative performance of our classifiers, we performed two sets of experiments. First, for each fMRI study we analyzed the performance of GNB, linear SVM, and kNN (with $k \in \{1, 3, 5, 7, 9\}$) using as input to the classifier *all* voxels in the ROIs selected for those studies. Here the performance metric is classification error, except for the Semantic Categories study where the metric is normalized rank error. The performance reported for a specific study is the mean error over all single-subject classifiers trained for that study, as obtained by leave-one-example-out-from-each-class cross validation. Because the Semantic Categories study is not a binary classification task, we did not experiment with SVMs on this specific study.

The results are shown in Table 1 in the rows indicating no feature selection. The table reports results for the better-performing variant of GNB in each study. In the Syntactic

Table 1. Error rates for classifiers across all studies.

Study	Examples per class	Feature selection	GNB	SVM	1NN	3NN	5NN	9NN
Picture vs.	40	Yes	0.18	0.11	0.22	0.18	0.18	0.19
Sentence	40	No	0.34	0.34	0.44	0.44	0.41	0.38
Semantic	32	Yes	0.08	N/A	0.31	0.21	0.17	0.14
Categories	32	No	0.10	N/A	0.40	0.40	0.40	0.25
Syntactic	10	Yes	0.25	0.28*	0.39	0.39	0.38	0.34
Ambiguity	10	No	0.41	0.38	0.50	0.46	0.47	0.43

Each table entry indicates the mean test error averaged over all single-subject classifiers trained for a particular fMRI study and learning method. The rows with Feature Selection "No" show results when using all voxels within the available ROIs. The rows with Feature Selection "Yes" show results of the feature selection method that produced the lowest errors. In every case except one, this was the "Active" feature selection method described in Section 6.3.1. The exception is the entry marked with the "*", for which "RoiActive" feature selection worked best. The variant of GNB which produced the strongest results (and which is therefore reported in this table) is GNB-SharedVariance for the Syntactic Ambiguity study, and GNB-DistinctVariance for the other two studies.

Ambiguity study this was GNB-SharedVariance, and in the other two studies it was GNB-DistinctVariance.

As can be seen in the table, the GNB and SVM classifiers outperformed kNN. Examining the performance of kNN, one can also see a trend that performance generally improves with increasing values of k.

Our second set of experiments examined the performance of the classifiers when used in conjunction with feature selection. The specific feature selection methods we considered are described in detail in the next subsection. For each study and learning method the table reports results using the most successful feature selection method, in the table row indicating feature selection "Yes." In all cases except one (the table entry marked by the "*"), the most successful feature selection method was the "Active" method described in Section 6.3.1.

As can be seen in Table 1, performing feature selection produced a large and consistent improvement in classification error across all studies and learning methods. As in the experiments with no feature selection, GNB and SVM outperform kNN when feature selection is used, and the performance of kNN improves as k increases.

6.2.1. Analysis. One clear trend in this data is that kNN fared less well than GNB or SVMs across all studies and conditions. In retrospect, this is not too surprising given the high dimensional, sparse training data sets. It is well known that the kNN classifier is sensitive to irrelevant features, as these features add in irrelevant ways to the distance between train and test examples (Mitchell, 1997). This explanation for the poor performance of kNN is also consistent with the dramatic improvement in kNN performance resulting from feature selection. As the table results indicate, feature selection sometimes reduces kNN error by a factor of two or more, presumably by removing many of these irrelevant, misleading features.

As discussed in Section 4.2, the two variants of GNB we considered differ only in the number of distinct parameters estimated when modeling variances in the class conditional

distributions of voxel activities. GNB-SharedVariance estimates a single variance independent of the class, whereas GNB-DistinctVariance estimates a distinct variance per class. We found that GNB-SharedVariance performed better in the study containing the fewest examples per class (Syntactic Ambiguity), whereas GNB-DistinctVariance performed better in the other studies. This empirical result is consistent with a general bias-variance tradeoff: pooling data in order to estimate fewer parameters generally leads to lower variance estimates, but to higher bias. Thus, in general as the number of available examples increases, we expect GNB-DistinctVariance to outperform GNB-SharedVariance, all other things being equal.

Notice that the SVM outperformed GNB in the Picture vs. Sentence study for which there were 40 examples per class, but not in the Semantic Categories study for which there were only 10 examples per class. While this may be due to various factors, it is interesting to observe that this trend is consistent with recent results of Ng and Jordan (2002), in which they provide empirical and theoretical arguments that GNB often outperforms Logistic Regression when data is scarce, but not when data becomes more plentiful. They explain this by pointing out that although Logistic Regression asymptotically (in the number of training examples) outperforms GNB, Logistic Regression requires $O(n)$ examples to reach its asymptote while GNB requires only $O(\log n)$, where n is the number of features. In our case we note that SVM, like Logistic Regression, requires $O(n)$ examples, in comparison to the $O(logn)$ required by GNB. While our empirical results are too sparse to prove a statistically significant trend for the relative performance of SVM and GNB, it is nevertheless interesting to note our results are consistent with the analysis of Ng and Jordan (2002).

In summary, we found when training fMRI classifiers across a variety of data sets and target functions that GNB and SVM outperformed kNN quite consistently. Furthermore, feature selection has a large and consistent beneficial impact across all studies and learning methods.

6.3. Which feature selection method works best?

Given that our classification problem involves very high dimensional, noisy, sparse training data, it is natural to consider feature selection methods to reduce the dimensionality of the data before training the classifier. As we discussed in the previous section, and as summarized in Table 1, feature selection leads to large and statistically significant improvements in classification error across all three of our case studies. This section discusses in detail the feature selection methods explored in our work, and some surprising lessons learned regarding which feature selection methods worked best.

6.3.1. Approach. Within the field of Machine Learning, the most common approach to feature selection when training classifiers is to select those features that best *discriminate* the target classes. For example, given the goal of learning a target classification function $f : X \rightarrow Y$, one common approach to feature selection is to rank order the features of X by their mutual information with respect to the class variable Y, then to greedily select the n highest scoring features.

Given the nature of classification problems in the fMRI domain (and a variety of other domains as well), a second general approach to feature selection is also possible. To illustrate, consider the problem of learning a Boolean classifier $f : X \to Y$ where $Y = \{1, 2\}$, given training examples labeled as belonging to either class 1 or class 2 (e.g., learning to distinguish whether the subject is viewing a picture or sentence). In fMRI studies, we naturally obtain *three* classes of data rather than two. In addition to data representing class 1 and class 2, we also obtain data corresponding to a third "fixation" or "rest" condition. This fixation condition contains data observed during the time intervals between trials, during which the subject is generally at rest (e.g., they are examining neither a picture nor a sentence, but are instead staring at a fixation point). Thus, we can view the data associated with class 1 and class 2 as containing some signal conditioned on the class variable Y, whereas the data associated with fixation contains no such signal, and instead contains only background noise relative to our classification problem. In this setting, we can consider a second general approach to feature selection: score each feature by how well it discriminates the class 1 or class 2 data from this zero signal data. In the terminology of fMRI, we score each feature based on how *active* it is during the class 1 or class 2 intervals, relative to the fixation intervals. The intuition behind this feature selection method is that it emphasizes choosing voxels with large signal-to-noise ratios, though it ignores whether the feature actually distinguishes the target classes.

We refer to this general setting as the "zero signal" learning setting, summarized in figure 3. Notice many classification problems involving sensor data can be modeled in

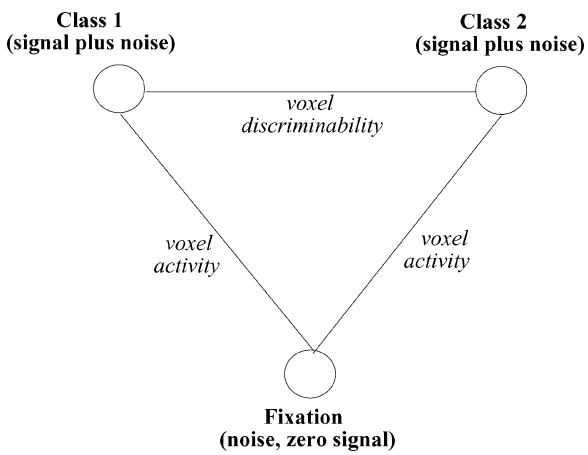

Figure 3. The "zero signal" learning setting. Boolean classification problems in the fMRI domain naturally give rise to three types of data: data corresponding to the two target classes plus data collected when the subject is in the "fixation" or "rest" condition. We assume the data from class 1 and class 2 are composed of some underlying signal plus noise, whereas data from the fixation condition contains no relevant signal but only noise. In such settings, feature selection methods can consider both *voxel discriminability* (how well the feature distinguishes class 1 from class 2), and *voxel activity* (how well the feature distinguishes class 1 or class 2 from the zero signal class).

terms of this zero signal learning setting (e.g., classifying which of two people is speaking based on voice data, where the zero signal condition corresponds to background noise when neither person is speaking). Therefore, understanding how to perform feature selection and classification within this setting has relevance beyond the domain of fMRI. In fact, within the fMRI literature it is common to use activity to select a subset of relevant voxels, and then to compare the behavior of this selected subset over various conditions.

In the experiments summarized below, we consider feature selection methods that select voxels based on both their ability to distinguish the target classes from one another (which we call *discriminability*), and on their ability to distinguish the target classes from the fixation condition (which we call *activity*). Although each feature consists of the value of a single voxel at a single time, we group the features involving the same voxel together for the purpose of feature selection, and thus focus on selecting a subset of voxels. In greater detail, the feature selection (voxel selection) methods we consider here are:

- *Select the n most discriminating voxels* (*Discrim*). In this method, a separate classifier is trained for each voxel, using only the observed fMRI data associated with that voxel. The accuracy of this single-voxel classifier over the training data is taken as a measure of the discriminating power of the voxel, and the n voxels that score highest according to this measure are selected. Note when reporting cross validation errors on final classifiers using this feature selection method, features are selected separately for each cross-validation fold in order to avoid using data from the test fold during the feature selection process. Thus, the voxels selected may vary from fold to fold.
- *Select the n most active voxels* (*Active*). In this method, voxels are selected based on their ability to distinguish either target class from the fixation condition. More specifically, for each voxel, v, and each target class y_i, a t-test is applied to compare the voxel's fMRI activity in examples belonging to class y_i to its activity in examples belonging to fixation periods. The first voxels are then selected by choosing for each target class y_i the voxel with the greatest t statistic. The next voxels are selected by picking the second most active voxel for each class, and so on, until n voxels are chosen. Notice the selected voxels may distinguish just one target class from fixation, or may distinguish both target classes from fixation.
- *Select the n most active voxels per Region of Interest* (*roiActive*). This is similar to the Active method above, but ensures that voxels are selected uniformly from all regions of interest (ROIs) within the brain. More precisely, given m prespecified ROIs, this method applies the Active method to each ROI, selecting n/m voxels from each. The union of these voxel sets is returned as the set of n selected voxels.

The approaches above for selecting voxels can be combined with methods for averaging the values of multiple features (in space or time), and with methods that select data over a sub-interval in time. We experimented with various combinations of such approaches, and report here on the above three methods (Discrim, Active, roiActive) as well as a fourth method derived from roiActive:

- *Calculate the mean of active voxels per ROI* (*roiActiveAvg*). This method first selects n/m voxels for each of the m ROIs using the roiActive method. It then creates a single

"supervoxel" for each ROI, whose activity at time t is the mean activity of the selected ROI voxels at time t.

6.3.2. Results. We experimented with each of these four feature selection methods, over each of the three case study data sets. For comparison purposes we also considered using all available features (denoted *"AllFeatures"* in our results tables).

Tables 2–4 present summarized results of feature selection experiments for each of the three fMRI studies. Each table shows the best errors obtained for each feature selection

Table 2. Picture vs. Sentence study—GNB classifier errors by subject and feature selection method.

Feature selection	Average error	A	B	C	D	E	F	G	H	I	J	K	L	M
AllFeatures(~10,000)	0.34	0.50	0.14	0.40	0.11	0.35	0.45	0.47	0.32	0.31	0.35	0.29	0.46	0.21
Discrim(1440)	0.32	0.39	0.09	0.39	0.06	0.27	0.39	0.39	0.30	0.34	0.39	0.36	0.55	0.21
Active(240)	0.18	0.29	0.09	0.24	0.04	0.15	0.34	0.29	0.05	0.19	0.10	0.15	0.35	0.10
roiActive(240)	0.23	0.37	0.12	0.26	0.16	0.19	0.39	0.31	0.12	0.25	0.20	0.20	0.32	0.15
roiActiveAvg(120)	0.27	0.39	0.15	0.36	0.12	0.22	0.39	0.30	0.22	0.24	0.29	0.26	0.45	0.17

The first column indicates the feature selection method, along with the number of features selected ("AllFeatures" indicates using all available features, which varies by subject). The second column indicates the average error over all 13 single-subject classifiers when using the feature selection method. Remaining columns indicate errors for individual subjects A through M. For each method, the number of features was chosen to minimize the average error over all subjects.

Table 3. Syntactic ambiguity study—GNB classifier errors by subject and feature selection method.

Feature selection	Average error	A	B	C	D	E
All(~2500)	0.41	0.25	0.55	0.50	0.30	0.45
Discrim(80)	0.38	0.20	0.60	0.45	0.15	0.50
Active(4)	0.25	0.20	0.25	0.40	0.25	0.15
roiActive(20)	0.27	0.30	0.35	0.30	0.20	0.20
roiActiveAvg(160)	0.35	0.25	0.40	0.30	0.35	0.45

Results are presented using the same format as Table 2.

Table 4. Semantic categories—GNB errors by subject and feature selection method.

Feature selection	Average error	A	B	C	D	E	F	G	H	I	J
All(~10,000)	0.100	0.13	0.17	0.04	0.12	0.06	0.07	0.20	0.04	0.14	0.05
Discrim(3200)	0.100	0.11	0.18	0.05	0.12	0.06	0.07	0.19	0.04	0.14	0.05
Active(2000)	0.083	0.11	0.12	0.04	0.10	0.06	0.07	0.11	0.04	0.13	0.05
roiActive(2400)	0.087	0.12	0.13	0.04	0.11	0.06	0.07	0.12	0.04	0.14	0.04

Results are presented using the same format as Table 2.

method and for each subject considered in the study, when using a GNB classifier. Here the best error refers to the lowest error achieved by varying the number of selected voxels.

The optimal number of voxels selected varied by study and feature selection method, typically ranging from 1 to 30% of the total number available within the selected ROIs, hence we focused experiments within this range. For each feature selection method and study, the number of features was chosen to minimize the mean error of all single-subject classifiers trained for this study. This number is shown in parentheses next to the feature selection method in the table. The specific numbers of features considered were 120, 240, 480, 960, 1440, 1920, 2400, 2880, and 3360 for the Picture versus Sentence study, 100, 200, 400, 800, 1200, 1600, 2000, 2400, 2800, 3200, and 3600 for the Semantic Categories study, and 4, 20, 40, 80, 160 for the Syntactic Ambiguity study.

These results indicate that using feature selection leads to improved classifier error in all three studies, and that all of our feature selection methods improve over no feature selection in the vast majority of cases. For example, the Active feature selection method outperforms the approach of using all features in 23 of the 28 single-subject classifiers trained over the three studies, and yields equivalent performance in the remaining 5 cases.

A second strong trend in the results is the dominance of feature selection methods based on activity (Active, roiActive, roiActiveAvg) over those based on discriminability (Discrim). As can be seen in the tables, *every* activity-based feature selection method outperforms the discriminability-based method, on every case study. Of these activity-based methods, the Active method yields best accuracy, and it outperforms the Discrim method in 20 of the 28 single-subject classifiers trained across the three studies, yielding inferior performance in only 1 of the 28 cases. Note also that the Active method selects substantially fewer features than Discrim in all three fMRI studies.

6.3.3. Analysis. It is at first surprising to observe that selecting features based on their activity level works dramatically better than selecting them based on their ability to discriminate the target classes. Given that the end goal is to discriminate the target classes, and that selecting features based on discriminability is the norm in machine learning applications, one might well expect discriminability to have been the dominant method. Below we look deeper into why we observe the opposite result in all three fMRI studies.

One situation in which we might expect activity-based feature selection to outperform discrimination-based methods is when data dimensionality is very high, noise levels are high, training data are sparse, and very few voxels contain a signal related to the target classes. In such cases, we should expect to find that some voxels that are truly irrelevant appear nonetheless to be good discriminators over the sparse sample of training data—even when using cross validation to test their discrimination power. The larger the set of such irrelevant voxels, the more likely that a feature selection strategy focused on discrimination would select such overfitting voxels, and be unable to distinguish these from truly informative discriminating voxels. However, in this same case we might expect that choosing voxels with high signal-to-noise ratios would be a useful strategy, as it would remove from consideration the large number of irrelevant voxels (i.e., those with no signal, but only noise). In fact, our activity-based feature selection strategies select exactly this kind of high signal-to-noise ratio voxels. The bottom line is that each feature selection strategy runs its own risk:

discrimination-based methods run the risk of selecting voxels that only coincidentally fit the noisy training sample, whereas activity-based methods run the risk of choosing high signal-to-noise voxels that cannot discriminate the target classes. Which risk is greater depends on the exact problem, but the relative risk for the discrimination-based method grows more quickly with increasing data dimension, increasing noise level, decreasing training set size, and an increasing fraction of irrelevant features.

To explore this conjecture, let us examine the actual characteristics of the voxels selected by these two methods in our data. In particular, we will focus on a single subject in the Semantic Categories study: subject G, whose best average normalized rank error (0.11) is obtained by a GNB classifier using 800 voxels chosen using the Active feature selection strategy. The Discrim method also obtains its best error (0.19) using 800 voxels for this particular subject. What is the difference in these two sets of voxels selected using these two methods? Are there in fact differences in the degree of overfitting between the two sets?

Let us consider three sets of voxels from subject G: those chosen by Active feature selection but not by Discrim ("ActiveOnly"), those chosen by Discrim by not by Active ("DiscriminatingOnly"), and those chosen independently by both methods ("Intersection"). For this particular subject, there are 251 voxels in the Intersection set, and 549 in each of ActiveOnly and DiscriminatingOnly. Training a GNB classifier using only the voxels in Intersection yields an error of (0.106), slightly but not significantly better than the error from the Active voxels.

Figure 4 shows the degree of overfitting for each of these three sets of voxels. On the left, panel (a) provides a scatterplot of training set error (horizontal axis) versus test set error (vertical axis). The straight line indicates where training error equals test error. Notice all three sets of voxels overfit to some degree (i.e., test error is generally greater than or equal to training error), but the cluster of voxels ranges furthest from the straight line for the DiscriminatingOnly voxels. On the right, panel (b) provides a histogram showing the number of voxels in each set that overfit to varying degrees. Note the DiscriminatingOnly voxel set contains many more voxels that overfit to a large degree. Based on the data summarized in this figure, it is clear that the degree of overfitting is indeed greater in this case for voxels selected by Discrim than those selected by Active. It is also clear that the voxels in Intersection suffer the least overfitting.

A different view into the character of these three voxel sets is provided by figure 5. The top portion of this figure plots the 10 voxels with the best training set error from each of the three sets. Each voxel plot shows the learned Gaussian model for each of the twelve target classes. Notice the greater spread of these models for the voxels chosen by the Discrim method (DiscriminatingOnly and Intersection) than for the ActiveOnly set. The bottom portion of figure 5 provides a scatter plot of standard deviation (horizontal axis) versus test error (vertical axis) for the three voxel sets. Notice the significantly lower standard deviation for the ActiveOnly set.

Above we suggested that the Discrim method for feature selection carries a risk of selecting voxels that overfit the data. The above data, especially from figure 4, indicates that in fact overfitting is greater for Discrim than for Active in our data. We also suggested the Active method carries the counterbalancing risk of selecting irrelevant voxels. Is this

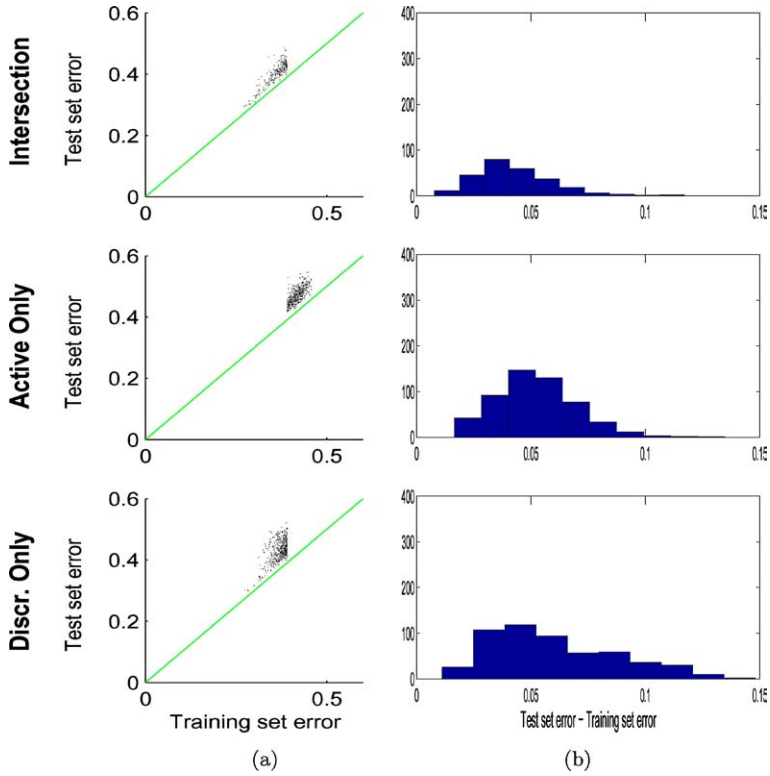

Figure 4. (a) Scatter plots of training set error (horizontal axis) against test set error (vertical axis) for the voxels in each subset (row). (b) Histograms depicting for each voxel subset the number of voxels that overfit to various degrees. The horizontal axis in this case is (test error minus training error), measuring the degree of overfitting.

in fact occurring in our case? The plots in figure 5 show that the ActiveOnly set of voxels does appear to contain voxels that are poor discriminators among the twelve target classes.

To understand the impact of poor discriminators (irrelevant voxels) selected by the Active method, consider the relative weight of the relevant versus irrelevant voxels used by a GNB. Given an instance $\vec{\mathbf{x}}$ to be classified, the log odds assigned by the GNB for two classes c_i and c_j is

$$\log \frac{\hat{P}(c_i \mid \vec{\mathbf{x}})}{\hat{P}(c_j \mid \vec{\mathbf{x}})} = \log \frac{\hat{P}(c_i)}{\hat{P}(c_j)} + \sum_k \log \frac{\hat{P}(x_k \mid c_i)}{\hat{P}(x_k \mid c_j)}$$

where x_k is the observed value for the kth feature of $\vec{\mathbf{x}}$, and where $\hat{P}$ denotes distributions estimated by GNB based on the training data. Note the GNB classifier will predict class c_i if the above log odds ratio is positive, and c_j if it is negative. Thus, the decision of the GNB is determined by a linear sum, where each voxel contributes one term to the sum.

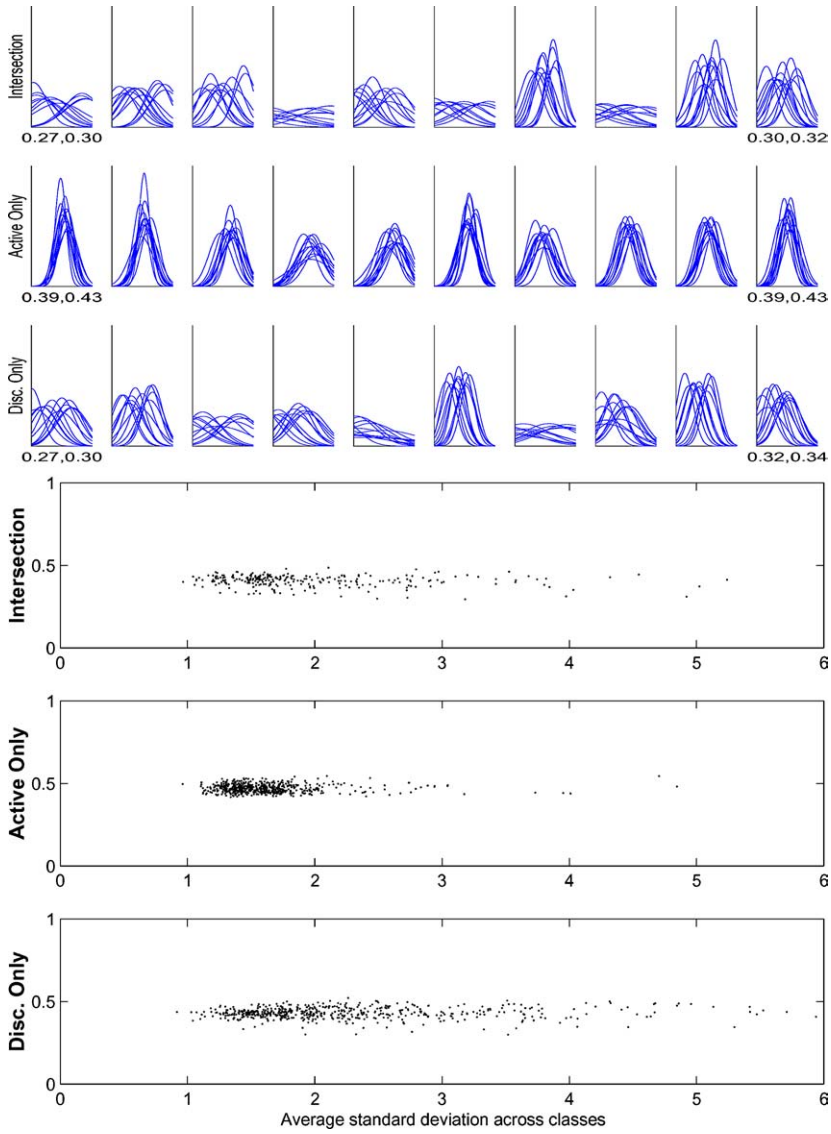

Figure 5. In the top half of the figure, each plot shows the 12 learned class probability densities $\hat{P}(x_k \mid c_i)$ for a single voxel x_k, for subject G in the Semantic Categories study. The x axis ranges from -5 to 5. Each row contains the 10 voxels with the lowest training set errors from each voxel subset, sorted by increasing error. For reference, the leftmost and rightmost plots in each row have their (training set, test set) error values below them. The bottom half of the figure provides scatterplots depicting the average class standard deviation (horizontal axis) against the test error (vertical axis) for each voxel subset (row). Note the higher variance for the Discriminating voxels (Intersection and DiscriminatingOnly).

Now let us consider a voxel x_k which is truly irrelevant to the classification (i.e., where the true distributions $P(x_k \mid c_i)$ and $P(x_k \mid c_j)$ are identical). First, consider the situation in which the learned estimates $\hat{P}(x_k \mid c_i)$ and $\hat{P}(x_k \mid c_j)$ are also identical. In this case the fraction involving x_k will be equal to 1, its log will be equal to 0 regardless of the observed value of x_k, and voxel x_k will therefore have no influence on the final GNB decision. Now consider the situation in which $\hat{P}(x_k \mid c_i)$ and $\hat{P}(x_k \mid c_j)$ differ (e.g., due to overfitting) despite the fact that $P(x_k \mid c_i) = P(x_k \mid c_j)$. In this case, the x_k term will in fact be non-zero, and will have a detrimental, randomizing influence on the final GNB classification. Is this in fact occurring in our data? The plots in figure 5 suggest that the Active voxels that are irrelevant (i.e., those in ActiveOnly) do indeed have strongly overlapping $\hat{P}(x_k \mid c_i)$ distributions, limiting the magnitude of their contribution to the final GNB classification.

To summarize, we find clear empirical evidence that feature selection consistently improves classification accuracy in our domain. Furthermore, we find clear empirical evidence that activity-based feature selection methods consistently outperform discrimination-based methods (*every* activity-based feature selection method we considered outperformed discrimination-based feature selection, in each of our three fMRI studies). The general characteristic of our problem domain that enables this kind of feature selection is summarized in figure 3. In particular, the key characteristic is the availability of a third category of data in which neither of the target classes is represented—data which we refer to as the "zero signal" class of data. We conjecture, and support with a variety of empirical observations, that activity-based feature selection may outperform discrimination-based feature selection in zero-signal classification problems, especially with increasing data dimension, noise, and data sparsity, and as the proportion of truly relevant features decreases. Given that a variety of sensor-based classification problems fit this zero-signal learning setting, we believe one of the most significant lessons learned from the fMRI domain is the utility of activity-based feature selection in such domains. Examples of such sensor-based classification problems include speaker voice identification (where zero signal data corresponds to background microphone noise when nobody is speaking), and video object classification in fixed-camera settings (where zero signal data corresponds to background images containing no objects of interest). Additional research is needed to formalize this zero-signal learning setting and explore its relevance in these and other domains. One especially promising direction for future work is to understand how best to blend activity-based and discrimination-based feature selection to optimize learning accuracy.

6.4. Can one train classifiers across multiple subjects?

All results discussed so far in this paper have focused on the problem of training subject-specific classifiers. This section considers the question of whether it is possible to train classifiers that apply across multiple human subjects, including human subjects who are not part of the training set. This is an important goal because if it is feasible it opens the possibility of discovering subject-independent regularities in brain activity associated with different types of cognitive processing, it opens the possibility of sharing trained classifiers among researchers analyzing fMRI data collected from many different people, and it makes

it possible to pool training data from multiple subjects to alleviate the sparse data problem in this domain.

The biggest obstacle to analysis of multiple-subject fMRI data is anatomical variability among subjects. Different brains vary significantly in their shapes and sizes, as is apparent from the images taken from three different brains in figure 2. This variability makes it problematic to register the many thousands of voxels in one brain to their precise corresponding locations in a different brain. One common approach to this problem is to transform (geometrically morph) fMRI data from different subjects into some standard anatomical space, such as Talairach coordinates (Talairach & Tournoux, 1988). The drawback of this method is that the morphing transformation typically introduces an error on the order of a centimeter, in aligning the millimeter-scale voxels from different brains. The alternative that we consider here is to instead abstract the fMRI activation, using the mean activation within each ROI as the classifier input, instead of using individual voxel activations. In other words, we treat each ROI as a large "supervoxel" whose activation is defined by the mean activation over all the voxels it contains. Note this is equivalent to using our roiActiveAverage feature selection method, including all voxels within each ROI. Despite the anatomical variability in brain sizes and shapes, it is relatively easy for trained experts to manually identify the same set of ROIs in each subject.

A second difficulty that arises when training multiple-subject classifiers is that the intensity of fMRI response to a particular stimulus is often different across subjects. To partially address this issue, we employ a normalization method that linearly rescales the data from different subjects into the same maximum-minimum range. While there are many cross-subject differences that cannot be addressed by this simple linear transformation, we have found this normalization to be useful. We have also found that a similar normalization method can sometimes reduce classification error for single-subject classifiers, when used to normalize data across different trials associated with a single subject.

Of course a third difficulty that arises is simply that different people may think differently, and we have no reason to assume a priori that we would find the same spatial-temporal patterns of fMRI activation in two subjects even if we could perfectly align their brains spatially, and perfectly normalize the relative intensities of their fMRI readings. Thus, the question of whether one can train multiple-subject classifiers is partly a question of whether different brains behave sufficiently similarly to exhibit a common pattern of activation.

We performed experiments to train multiple-subject classifiers using two data sets: the Picture versus Sentence data, and the Syntactic Ambiguity data. The following two subsections describe these two experiments in turn. We did not attempt to train multiple-subject classifiers for the third dataset, Semantic Categories, because we expected the detailed patterns of activity that distinguish word categories would be undetectable once the voxel-level data was abstracted to mean ROI activity.

6.4.1. Sentence versus picture study. We trained multiple-subject classifiers for the Sentence versus Picture study, to discriminate whether the subject was viewing a picture or a sentence. Multiple-subject classifiers were trained using data from 12 of the 13 subjects, abstracting the data from each subject into seven ROI supervoxels. To evaluate the error of these trained classifiers, we used leave-one-subject-out cross validation. In particular, for

Table 5. Errors for multiple-subject classifier, sentence versus picture study.

Classifier	Leave-1-subject-out error
GNB	0.30 ± 0.028
SVM	0.25 ± 0.026
1NN	0.36 ± 0.029
3NN	0.33 ± 0.029
5NN	0.32 ± 0.028

The right column shows the error of a multi-subject classifier when applied to subjects withheld from the training set, using leave-one-subject-out cross validation. Results are obtained using normalization and 7 ROIs. All classifiers are trained by averaging all voxels in an ROI into a supervoxel. 95% confidence intervals are computed under the assumption that test examples are identically, independently Bernoulli distributed. The error of a random classifier is 0.50. In this analysis we employed the GNB-DistinctVariance version of GNB.

each subject we trained on the remaining 12 subjects, measured the error on this held out subject, then calculated the mean error over all held out subjects. Notice that in this case there are 960 examples available to train this multiple-subject classifier, in contrast to the 80 examples available to train the single-subject classifiers described earlier.

The results, summarized in Table 5, show that the linear SVM learns a multiple-subject classifier that achieves error of 0.25 ± 0.026 over the left out subject. This is highly statistically significant compared to the 0.50 error expected of a default classifier guessing the majority class. This indicates that it is indeed possible to train a classifier to capture significant subject-independent regularities in brain activity that are sufficiently strong to detect cognitive states in human subjects who are not part of the training set. As in earlier experiments, we note that SVM and GNB again outperform kNN, and that the performance of kNN improves with increasing values of k.

Although these results demonstrate that it is possible to learn multiple-subject classifiers with accuracies better than random, the accuracies are below those achieved by the single-subject classifiers summarized in Table 1. For example, the multiple-subject SVM classifier error of 0.25 is less accurate than the mean single-subject classifier error of 0.11 reported in Table 1. This difference could be due to a variety of factors, ranging from the lower spatial resolution encoding of fMRI inputs in the multiple-subject classifier (i.e., using supervoxels instead of millimeter-scale voxels), to possible differences in brain activation over different subjects.

In a second set of experiments we directly compared training single-subject versus multiple-subject classifiers, this time using identical training methods and identical encodings for the classifier input (roiActiveAverage, using all voxels within seven manually selected ROIs). In these experiments we used the same Picture versus Sentence data, but

Table 6. Errors for single-subject and multiple-subject classifiers, when trained on P-then-S, and S-then-P data.

Data set	Classifier	Single-subject classifier	Multiple-subject classifier
S-then-P	GNB	0.10 ± 0.024	0.14 ± 0.030
S-then-P	SVM	0.11 ± 0.025	0.13 ± 0.029
S-then-P	1NN	0.13 ± 0.028	0.15 ± 0.031
S-then-P	3NN	0.12 ± 0.027	0.13 ± 0.029
S-then-P	5NN	0.10 ± 0.025	0.11 ± 0.027
P-then-S	GNB	0.20 ± 0.033	0.20 ± 0.034
P-then-S	SVM	0.17 ± 0.031	0.22 ± 0.036
P-then-S	1NN	0.38 ± 0.041	0.26 ± 0.038
P-then-S	3NN	0.31 ± 0.039	0.24 ± 0.037
P-then-S	5NN	0.26 ± 0.037	0.21 ± 0.035

The third column shows the average error of classifiers trained for single subjects. The fourth column shows the error of multi-subject classifiers applied to subjects withheld from the training set. Results are obtained using normalization. All classifiers are trained based upon averaging all voxels in an ROI into a supervoxel. 95% confidence intervals are computed under the assumption that test examples are i.i.d. Bernoulli distributed. The error of a random classifier is 0.50.

this time we partitioned the data into two disjoint subsets: trials in which the sentence was presented before the picture (which we will refer to as S-then-P), and trials in which the picture was presented before the sentence (which we will call P-then-S). Notice that separating the data in this fashion results in an easier classification problem, because all examples of one stimulus (e.g., sentences) occur in the same temporal context (e.g., following only the rest period in the S-then-P dataset, or following only the picture stimulus in the P-then-S dataset), and hence exhibit less variability. For each of these two data subsets we trained both multiple-subject and single-subject classifiers, using GNB, SVM, and kNN classifiers, and employing roiActiveAverage feature selection with all voxels in the selected ROIs. The results are summarized in Table 6. Note for comparison we present both the leave-1-subject-out error of the multiple-subject classifiers, and the average leave-one-example-per-class-out error of the corresponding single-subject classifiers.

As in the first experiment, the multiple-subject classifiers achieve accuracies significantly greater than the 0.5 expected from random guessing. Interestingly, the multiple-subject classifiers achieve error rates comparable to those of the single-subject classifiers, and in a few cases achieve error rates superior to those of the single-subject classifiers—despite the fact that the multiple-subject classifiers are being evaluated on subjects outside the training set. Presumably this better performance by the multiple-subject classifier can be explained by the fact that it is trained using an order of magnitude more training examples, from twelve subjects rather than one. We interpret these results as strong support for the feasibility of training high accuracy classifiers that apply to novel human subjects. Note these classifiers are generally more accurate than those in the first experiment, presumably due to the easier classification task as discussed above.

6.4.2. Syntactic ambiguity study. We also attempted to train multiple-subject classifiers for the Syntactic Ambiguity study, to discriminate whether the subject was reading an ambiguous or non-ambiguous sentence. In this case, the error of the multi-subject classifier was 0.36 ± 0.094 under leave-one-subject-out cross validation, and correspondingly the average error of single-subject classifiers is 0.35 ± 0.092. The setting which produced this result was using the GNB classifier, minimum-maximum normalization, and the feature selection method roiActiveAvg, averaging the 20 most active voxels from each of four predefined ROI's (left and right Brocca, and left and right temporal regions) into supervoxels. These errors are significantly better than expected from a random classifier, 0.50. Unlike the Sentence versus Picture study, however, these results are quite sensitive to the particular selection of learning method and feature selection. Although we cannot draw strong conclusions from this result, it does provide modest additional support for the feasibility of training multiple-subject classifiers.

7. Summary and conclusions

We have presented results from three different fMRI studies demonstrating the feasibility of training classifiers to distinguish a variety of cognitive states, based on single-interval fMRI observations. This problem is interesting both because of its relevance to studying human cognition, and as a case study of machine learning in high dimensional, noisy, sparse data settings.

Our comparison of classifiers indicates that Gaussian Naive Bayes (GNB) and linear Support Vector Machine (SVM) classifiers outperform k Nearest Neighbor across all three studies, and that feature selection methods consistently improve classification error in all three studies. In comparing GNB to SVM, we found trends consistent with the observations in Ng and Jordan (2002), that the relative performance of generative versus discriminative classifiers depends in a predictable fashion on the number of training examples and data dimension. In particular, our experiments are consistent with the hypothesis that the accuracy of SVM's increases relatively more quickly than the accuracy of GNB as the data dimension is reduced via feature selection, and as the number of training examples increases.

Feature selection is an important aspect in the design of classifiers for high dimensional, sparse, noisy data. We defined a new classifier setting (the zero signal setting) that captures an important aspect of our fMRI classification problem, as well as a variety of other classification problems involving sensor data. In this setting, the available data includes not only examples of the classes to be discriminated (e.g., data when the subject is viewing a picture or a sentence), but also a class of "zero signal" data (e.g., when the subject is viewing neither a picture nor a sentence, but is simply fixating on the screen). Our experiments show that within our domain activity-based feature selection methods which take advantage of this zero signal data consistently outperform traditional discrimination-based feature selection methods that use only data from the target classes. Our data and our analysis also suggest that the relative benefit of activity-based versus discrimination-based feature selection will increase as data becomes more sparse, more noisy, higher dimensional, and as the fraction of relevant features decreases. As is clear from our description of the fMRI data sets, this domain represents a fairly extreme point along all four of these dimensions. We plan further

research to develop a more precise formal model of this zero signal setting, and to develop and experiment with feature selection strategies tuned to take maximal advantage of this setting.

In addition to training classifiers to detect cognitive states in single subjects, we also explored the feasibility of training cross-subject classifiers to make predictions across multiple human subjects. In this case, we found it useful to abstract the fMRI data by using the mean fMRI activity in each of several anatomically defined brain regions. Using this approach, it was possible to train classifiers to distinguish, e.g., whether the subject was viewing a picture or a sentence describing a picture, and to apply these successfully to subjects outside the training set. In some cases, the classification accuracy for subjects outside the training set equalled the accuracy achieved by training on data from just this single subject. Given this success in training multiple-subject classifiers, we plan additional research to explore a number of alternative approaches to cross-subject classification (e.g., instead of abstracting the data for each subject, map the different brain structures to a standard coordinate system such as Talairach coordinates).

There are many additional opportunities for machine learning research in the context of fMRI data analysis. For example, it would be useful to learn models of temporal behavior, in contrast to the work reported here which considers only data at a single time or time interval. Machine learning methods such as Hidden Markov Models and Dynamic Bayesian Networks appear relevant to this problem. A second research direction is to develop learning methods that take advantage of data from multiple studies, in contrast to the single study efforts described here. In our own lab, for example, we have accumulated fMRI data from over 800 human subjects. In order to develop learning methods to take advantage of such data, it will be necessary to address both how to combine data from multiple subjects and how to combine data from subjects presented with differing stimuli. A third research topic is to develop machine learning methods that could take as a starting point computational cognitive models of human processing, such as ACT-R (Anderson et al., 2004) and 4CAPS (Just, Carpenter, & Varma, 1999), using these as prior knowledge for guiding the analysis of fMRI data, and automatically refining these models to better fit observed experimental results.

As with many real-world machine learning case studies, our exploration of the fMRI problem domain has drawn on lessons learned from previous research in machine learning, and has yielded new lessons of its own. Given that fMRI is a problem involving very high dimensional, sparse data sets, we drew heavily on previously learned lessons from similar domains such as text classification. This led us to employ SVM, GNB, and kNN classification algorithms that have previously proven useful in such domains, and led us to aggressively explore feature selection methods for reducing the dimension of the data. The most significant new insight about learning to arise from our fMRI studies thus far is the identification of the zero-signal learning setting and the development of new and highly effective feature selection methods for this setting. In particular, our discovery of the unexpected dominance of activity-based feature selection methods over commonly used discrimination-based methods was an essential step toward training successful classifiers in this domain, and suggests that similar feature selection approaches may be useful in other high dimensional, sparse domains that fit the zero-signal learning setting. Looking

forward, we expect machine learning methods to have an increasing impact on the analysis of fMRI data as this field matures, and foresee additional opportunities for the fMRI domain to drive novel machine learning research, especially in problems related to discovery of representations for merging data from multiple subjects, and in learning temporal models of cognitive processes.

Acknowledgments

We are grateful to Luis J. Barrios for helpful discussions and detailed comments on various drafts of this paper. Thanks to Vladimir Cherkassky and Joel Welling for useful observations and suggestions during the course of this work, and to Paul Bennett for many helpful discussions and for writing part of the code used for the Semantic Categories study. We are also grateful for the detailed comments of two anonymous reviewers, which led to significant improvements to the final version of this paper.

Radu Stefan Niculescu was supported by a Graduate Fellowship from the Merck Computational Biology and Chemistry Program at Carnegie Mellon University established by the Merck Company Foundation and by National Science Foundation (NSF) grant no. CCR-0122581. Francisco Pereira was funded by the Center for Neural Basis of Cognition, a PRAXIS XXI scholarship from Fundação para a Ciência e Tecnologia, Portugal (III Quadro Comunitário de Apoio, comparticipado pelo Fundo Social Europeu) and a PhD scholarship from Fundação Calouste Gulbenkian, Portugal. Rebecca Hutchinson was supported by an NSF Graduate Fellowship. Support for collecting the fMRI data sets was provided by grant number N00014-01-1-0677 from the Multidisciplinary University Research Initiative (MURI) of the Office of the Secretary of Defense.

Notes

1. FIASCO is available at http://www.stat.cmu.edu/∼fiasco.
2. Notice this classification task is made more difficult by the fact that the first stimulus is always presented for four seconds, whereas the second stimulus is terminated as soon as the subject responds with a button press.
3. The experiment included four types of sentences. We consider here only two types, corresponding to the most and least ambiguous.

References

Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to building stimuli: Evidence and implications. *Neuron, 21*, 373–383.

Anderson, J. R. et al. (2004). An information-processing model of the BOLD response in symbol manipulation tasks. *Psychonomic Bulletin and Review* (in press).

Battig, W. F., & Montague, W. E. (1968). Category norms for verbal items in 56 categories: A replication and extension of the Connecticut norms. *Journal of Experimental Psychology Monograph, 80:3*, 1–46.

Blankertz, B., Curio, G., & Mller, K. R. (2002). Classifying single trial EEG: Towards brain computer interfacing. *Advances in Neural Inf. Proc. Systems (NIPS 2001), 14*, 157–164.

Bly, B. M. (2001). When you have a General Linear Hammer, every fMRI time-series looks like independent identically distributed nails. *Concepts and Methods in NeuroImaging Workshop*, NIPS.

Burges, C. (1998). A tutorial on support vector machines for pattern recognition. *Journal of data Mining and Knowledge Discovery, 2:2*, 121–167.

Caviness, V. S., Kennedy, D. N., Bates, J., & Makris, N. J. (1996). MRI-based parcellation of human neocortex: An anatomically specified method with estimate of reliability. *Journal of Cognitive Neuroscience, 8*, 566–588.

Chao, L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience, 2*, 913–919.

Chao, L., Weisberg, J., & Martin, A. (2002). Experience-dependent modulation of category-related cortical activity. *Cerebral Cortex, 12*, 545–551.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. Wiley and Sons.

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage, 19*, 261–270.

Eddy, W. et al. (1998). The challenge of functional magnetic resonance imaging. *Journal of Computational and Graphical Statistics, 8:3*, 545–558.

Friston, K. J. et al. (1995a). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping, 2*, 189–210,

Friston, K. J. et al. (1995b). Analysis of fMRI time-series revisited. *NeuroImage, 2*, 45–53.

Genovese, C. (1999). Statistical inference in functional magnetic resonance imaging. CMU Statistics Tech Report 674.

Goutte, C., Toft, P., Rostrup, E., Nielsen, F. A., & Hansen, L. K. (1998). On clustering fMRi time series. Technical Report IMM-REP-1998-11.

Haxby, J. et al. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*, 2425–2430, .

Hojen-Sorensen, P., Hansen, L. K., & Rasmussen, C. E. (1999). Bayesian modeling of fMRI time series. NIPS*99. Denver.

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed Representation of Objects in the Human Ventral Visual Pathway. *Proc. Nat. Acad. Sci. USA*, 96, 9379–9384, .

Joachims, T. (2001). A statistical learning model of text classification with support vector machines. In *Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR)*, ACM.

Just, M. A., Carpenter, P.A., & Varma, S. (1999). Computational modeling of high-level cognition and brain function. *Human Brain Mapping, 8*, 128–136.

Keller, T. A., Just, M. A., & Stenger, V. A. (2001). Reading span and the time-course of cortical activation in sentence-picture verification. *Annual Convention of the Psychonomic Society*, Orlando, FL.

Kjems, U., Hansen, L., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D., & Strother, S. C. (2002). The quantitative evalutation of functional neuroimaging experiments: Mutual information learning curves, *NeuroImage, 15*, 772–786.

Mason, R., Just, M., Keller, T., & Carpenter, P. (2004). Ambiguity in the Brain: What brain imaging reveals about the processing of syntactically ambiguous sentences, *Jounal of Experimental Psychology: Learning, Memory, and Cognition* (in press).

McKeown, M. J. et al. (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping, 6:3*, 160–188.

Menon, R. S., Luknowsky, D. L., & Gati, J. S. (1998). Mental chronometry using latency-resolved functional magnetic resonance imaging. *Proc. Natl. Acad. Sci (U.S.A.), 95*, 10902–10907.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mitchell, T. M., Hutchinson, R., Just, M., Niculescu, S. R., Pereira, F., & Wang, X. (2003). Classifying instantaneous cognitive states from fMRI data. In *Proceedings of the 2003 Americal Medical Informatics Association Annual Symposium*. Washington D.C.

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39*, 103–134.

Ng, A. Y., & Jordan, M. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. *Neural Information Processing Systems, 14*.

Penny, W. (2001). Mixture models with adaptive spatial priors. *Concepts and Methods in NeuroImaging* workshop at NIPS*01, Vancouver, British Columbia, Canada.

Rademacher, J., Galaburda, A. M., Kennedy, D. N., Filipek, P. A., & Caviness, V. S. (1992). Human cerebral cortex: Localization, parcellation, and morphometry with magnetic resonance imaging. *Journal of Cognitive Neuroscience, 4*, 352–374.

Strother S. C., Anderson, J., Hansen, L., Kjems, U., Kustra, R., Siditis, J., Frutiger, S., Muley, S., LaConte, S., & Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *Neuroimage, 15*, 747–771.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain*. Thieme, New York.

Wagner, A. D. et al. (1998). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science, 281*, 1188–1191.

Wang, X., Hutchinson, R., & Mitchell, T. M. (2003). Training fMRI classifiers to detect cognitive states across multiple human subjects. In *Proceedings of the 2003 Conference on Neural Information Processing Systems*, Vancouver.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Journal of Information Retrieval, 1:1/2*, 67–88.

Received April 30, 2003
Accepted April 8, 2004
Final manuscript May 14, 2004