Linear and Nonlinear Dimensionality Reduction in fMRI Data for Picture-Sentence Classification

Stuart Anderson
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
soa@ri.cmu.edu

Kevin Oishi
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
koishi@ri.cmu.edu

Abstract

fMRI data is represented in a space with very high dimensionality. Because of this, classifiers such as SVM and Naive Bayes may overfit this data. Dimensionality reduction methods are intended to extract features from data in a high dimensional space. Training a classifier on data in a lower dimension may improve the true error of the classifier beyond the performance obtained by training in a higher dimensional space. We experimented with the PCA[1], Isomap[2], and LLE[4] methods for extracting features from fMRI data of subjects viewing pictures and sentences. Additionally we experimented with an extension to the LLE algorithm, which improved classification.

1 Introduction

fMRI data is collected from human subjects as they perform tasks. The data is a three dimensional image of the activity in the brain segmented into 3mm voxels. There are approximately 5000 of these voxels, recorded every half second, in each brain image. We wish to reduce the dimensionality of this data for classification. We explored how different classification and reduction methods affected the classification accuracy.

Dimensionality reduction methods can be divided into two classes based on whether the projection onto a low dimensional manifold is linear or non-linear. One popular linear dimensionality method is Principle Component Analysis (PCA), a second linear method is using expert knowledge to eliminate some dimensions. Two nonlinear reduction methods we explored were Isomap[2] and Locally Linear Embedding (LLE)[4]. We also experimented with a small modification to LLE that improved classification performance. Two classification methods that have been used on fMRI data in the past are Linear Support Vector Machine (SVM) and Gaussian Nave Bayes (GNB) classification[6].

We investigated the effect of these dimensionality reduction methods on classification. We classified brain images taken from subjects either observing a picture or observing a sentence. The leave-one-out classification error rate for each classifier is reported.

2 Dimensionality reduction methods

2.1 Principle component analysis (PCA)

Principle Component Analysis (PCA) is a form of unsupervised feature selection. Given a set of k points in d-dimensional space $\{x^1, \ldots, x^k\}$, the PCA method finds the n-dimensional linear projection that minimizes the squared reconstruction error. This is equivalent to maximizing the variance of the data, and a greedy problem in the target dimension n. It can be shown that projecting onto the eigenvectors of the covariance matrix of the data with the k largest eigenvalues satisfies this condition. If the number of data points equals the dimension of the data, the eigenvectors of the covariance matrix are just the eigenvectors of matrix whose rows are the zero-mean centered data.[1] If the number of data points is not equal to the dimension of the data, we can still find the principle components without computing the full covariance matrix by finding the singular value decomposition of the matrix whose rows are the zero-mean centered data. In this case the eigengenes with the largest singular values correspond to the eigenvectors with the largest eigenvalues, and the eigenassay times the singular value matrix is the transformation of the data into principle component space. Given enough data, PCA is guaranteed to recover the true structure of data that lies on a linear projection of input state space.

2.2 Isomap

Isomap is another example of unsupervised feature selection. Isomap builds on Multi Dimensional Scaling (MDS), a generalization of PCA, but attempts to preserve the global structure through a non-Euclidean "geodesic" distance metric.[2] This approach finds non-linear degrees of freedom that describe the manifold embedded in the input space that the data lies on. The first step of Isomap is to build a neighborhood graph by connecting points either by computing k-NN at each point, or finding all points within some distance ϵ for each point. The next step is to estimate the geodesic distances between points using Dijkstra's algorithm on the neighborhood graph. We now use a form of MDS to compute the low-dimensional embedding by trying to preserve the ranking of the geodesic distance approximations. Isomap is also guaranteed to converge asymptotically to the true structure of the data manifold, provided the data comes from a distribution such that as the number of data points increases, the geodesic distance approximations become increasingly better.[3]

2.3 Locally Linear Embedding (LLE)

The key idea motivating LLE is the representation of local neighborhoods using a metric which is independent of the dimensionality of the space in which the local neighborhoods are embedded. Reduction is accomplished by representing the input data using this metric, then reconstructing the data in a lower dimension from the metric. Specifically, LLE uses the covariance matrix of the local neighborhood of a point as a dimensionality independent measure of the manifold surrounding that point. The eigenvectors corresponding to the largest eigenvalues of that covariance matrix form the basis of the tangent hyperplane of the low dimensional manifold at this point. This covariance matrix is used to compute a weighting vector such that the local point can be expressed as a weighted sum of its neighbors. These weighting vectors are then entered into a large sparse matrix relating the local neighborhoods of each point. An embedding is then computed by a global eigenvector solver which minimizes the reconstruction error of each point in the low dimensional space using the weighting vectors computed in the high dimensional space.[5]

2.4 Modified Locally Linear Embedding

We wanted to use the labels of our input data to improve the usefulness of LLE for classification. Our modification was to introduce an additional dimension to the input data with data points located along this dimension according to their classification. For each dataset we defined a hand-tuned parameter defining the distance between the two possible classifications, and placed unlabelled points half-way between the locations of each class. The effect of this modification is to alter the local neighborhoods of classified points but not unclassified points. This tends to produce an embedding in which points with different classifications are separable.

3 fMRI data set

Functional Magnetic Resonance Imaging (fMRI) measures the Blood Oxygen Level Dependent (BOLD) response in a brain, an indication of neural activity. We worked with fMRI data collected from 6 different subjects measured at a resolution of a few cubic millimeters (roughly 5000 voxels), collected every 500 milliseconds. Each subject participated in 54 trials. 14 of these trials were rest or fixation trials, where the subject either rested or fixated on a point on a screen. In the remaining 40 trials the subject was shown a series of two stimuli. The series was either a picture then a sentence, or a sentence then a picture. The trial consisted of showing the subject the first stimulus for 4 seconds, then a blank screen for 4 seconds, then the second stimulus for 4 seconds or until a button was pressed, then resting for 15 seconds. These trials were named for their first stimulus. Of the 40 trials, 20 were "Picture" trials and 20 were "Sentence" trials.

4 Procedure

4.1 fMRI preprocessing

The fMRI data used for these experiments was taken from subjects 04847, 04799, 05710, 04820, 05675, and 05680 of the database used by (Mitchell, 2004). We used the first 4 seconds of each trial as input data to the dimensionality reduction methods. During these 4 seconds the subject observed either a picture or a sentence. Since one image is acquired every half-second there are 8 images per trial and 320 images per subject over the course of 40 trials. The remaining data from each trial was discarded.

4.2 Feature selection and classification

For each subject we wished to establish the leave one out classification error rate of linear SVM and GNB classifiers as a function of the dimensionality reduction method used and the output dimensionality. Because PCA, LLE, and Isomap do not use the labeling of the data to produce embedding they were run only once for each subject. For all dimensions between 2 and the minimum of 100 and the largest dimension the method could produce we evaluated the leave one out error of Linear SVM and GNB. We used the linear SVM kernel provided by the svm_light package and the GNB implementation provided with fmri_core. For our modified version of LLE we needed to re-run the method for each point in the dataset, as the embedding depends on which point is unlabelled. We chose values for the tunable separation parameter of the modified LLE method by finding the minimum value that produced a linearly separable embedding of the points.

We were also given expert knowledge that the anatomical regions of interest (ROIs) CALC, LIPL, LT, LTRIA, LOPER, LIPS, or LDLPFC are strongly associated with the task of reading a sentence or looking at a picture. We were interested in the performance of our dimen-

sionality reduction and classification methods given this prior knowledge, and repeated this procedure after stripping the data of all voxels not labeled with these particular ROIs.

5 Results

Our main results are illustrated in figures 1-3. Figure 1 shows the leave-one-out error rate for picture-sentence classification using SVM on reduced data. In the limit, PCA outperforms LLE, Isomap, and our modified LLE. LLE performs moderately better than Isomap, and our modified LLE performs significantly better than LLE or Isomap, even beating PCA in low dimensions. Figure 2 compares the error rate of SVM on the reduction of expert-knowledge-masked data, and the error rate of SVM on the reduction of the full input data. This shows that for all three dimensionality reduction methods, masking the data with expert knowledge hurt classification accuracy. Figure 3 compares SVM to GNB using the same sets of reduced data. SVM produces much better results than GNB for PCA and LLE, but the classification accuracy of Isomap reduced data does not change significantly.

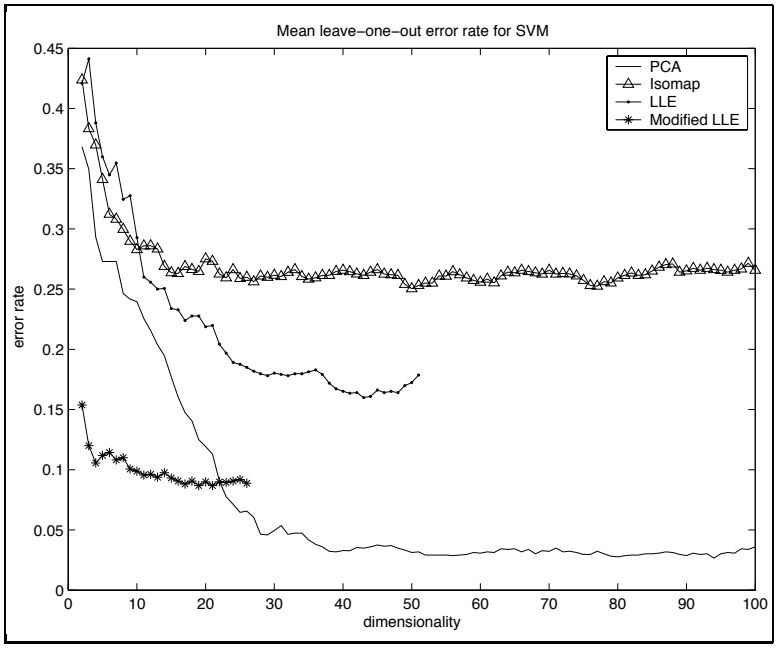

Figure 1: Mean leave-one-out error rate for picture-sentence classification using SVM on reduced data.

6 Discussion

6.1 Feature selection

One clear feature of our results is that classification on the PCA reduced data consistently outperformed classification on data reduced with the unlabelled nonlinear methods. We believe this may be related to the choices we made when preprocessing the data. Because we did not perform any time averaging of the data the noise in each sample is relatively large. If this noise is uniform over all dimensions it will tend to mask useful information in

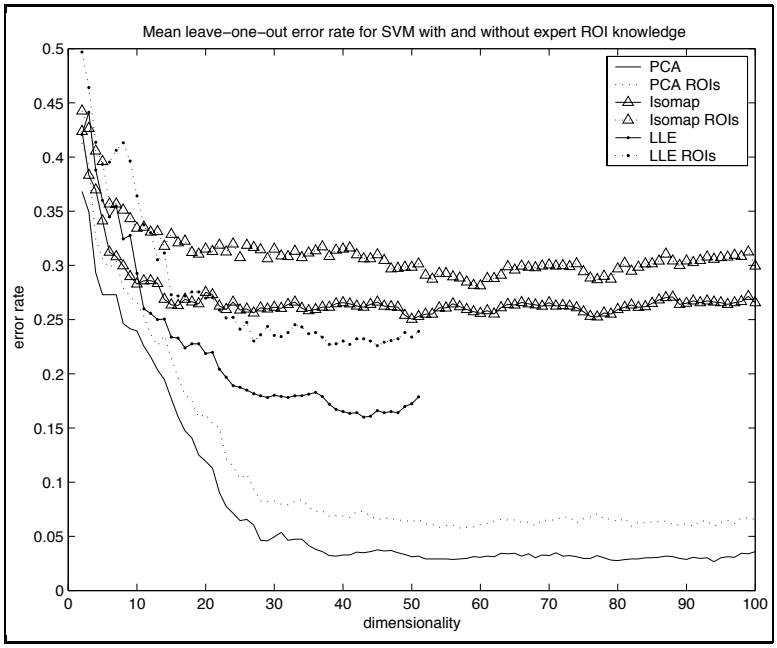

Figure 2: Mean leave-one-out error rate for picture-sentence classification using SVM on reduced data with and without and expert knowledge mask.

dimensions with low variance. Because PCA selects subspaces with high variance it will tend to select the subspaces with useful information. Another possible explanation for this result is that the voxels that are useful for classification are those that change value shortly after the stimulus is presented. A slight delay between stimulus and BOLD response is expected. This change would also give those dimensions a high variance and lead PCA to select them. In either case, we believe our results illustrate that PCA tends to select subspaces which are useful for classification, and that this tendency is a result of the underlying statistical structure of the data.

6.2 Classification methods

A second clear feature is linear SVM's lower leave one out error rate compared to GNB. The GNB classifier assumes that each component of the input vector is independent. This is clearly not the case in our data, where entire regions of the brain can become active at once. Because linear SVM does not make this same assumption we expect it to perform better on this dataset.

6.3 Regions of Interest (ROIs)

Additionally, we saw that removing data not in the regions of interest believed to be relevant to this classification problem increased the error rate of all classifiers for all dimensionality reduction methods. This result may initially seem unlikely, since one typically expects the addition of domain specific prior knowledge to improve classification accuracy. However, in this case the prior knowledge excludes a large portion of the available data from consideration. Our results show that when a dimensionality reduction method is left to decide which data is relevant without interference it produces an embedding more useful for classification that when it is initially assisted by this particular expert human knowledge. This

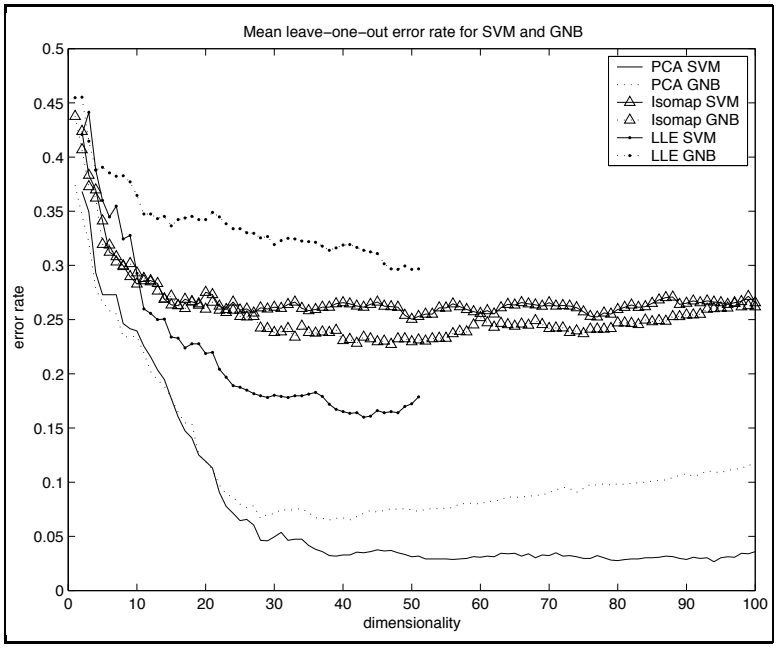

Figure 3: Mean leave-one-out error rate for picture-sentence classification using SVM and GNB on reduced data.

does not necessarily mean that a less restrictive form of this prior, a weighting, for example, would also fail to improve classification error rate.

6.4 Modified LLE

The modified LLE algorithm produced significantly better classification results at low dimensions than any other reduction algorithm, including PCA. We believe that using information about the classification of training points gives the method a significant advantage over the other methods we explored.

Consider an unclassified point that lies near two regions of the low dimensional manifold with opposite classifications. If the manifold is sampled sufficiently densely with sufficiently low noise then the classification of the nearest neighbor of the unclassified point is the best estimate of the point's classification, assuming that the point is generated from a Gaussian mixture distribution centered on the manifold. However, as the sampling density decreases or noise increases the nearest neighbor is no longer the best estimate of the nearest manifold. In this case we use the LLE embedding to estimate the manifold near the unclassified point then construct an embedding which reflects the point's relationship to the estimated manifold.

Suppose the noisy classified point closest to the unclassified point has the wrong classification. Then reconstruction error is minimized by moving the noisy point onto the manifold rather than moving all of its neighbors off of the manifold. Because the neighbors are shifted to better approximate the global structure of the manifold, the unclassified point is moved further from the incorrect manifold. Because of this effect we expect that k-nearest neighbor classification, as well as other classification methods, may have a lower error rate after the modified LLE method is applied to the data.

7 Conclusions

We explored the use of a variety of dimensionality reduction methods to assist the classification of fMRI data. We demonstrated that classifiers using the results of linear PCA outperform those using results from both the unlabeled nonlinear methods we tried, LLE and Isomap. Among the nonlinear methods LLE showed a modest improvement over Isomap. Additionally, we demonstrated a modification to LLE which used labeled training data to improve classification accuracy in low dimensions beyond that of any of the unlabeled methods.

We believe that part of the reason PCA performed well is related to the level of noise in the data. Preprocessing the data by time averaging or median filtering it might improve the relative performance of the nonlinear methods. Additionally, using the Birn Impulse Response function instead of direct BOLD measurements might improve classification accuracy. Finally, the expert knowledge about ROI importance might be made helpful to a classifier or dimensionality reduction method if it were used as a weighting rather than a mask.

An additional modification to LLE that could produce manifolds more closely reflecting the continuous changes in the subject's brains would be to build local neighborhoods from samples that were observed at consecutive or nearly consecutive times. This might do a better job of finding the paths through 'brain-space' that the subject's brain took during the trial and thereby improve classification accuracy by revealing important nonlinear features of the trial to the classifier.

A challenging extension to this classification problem would be to include fixation data in the trials as well as data from the second component of each trial in which the user was required to respond to the sentence. In our initial experiments with this portion of the data we found classification to be much more difficult.

References

[1] T.M. Mitchell. *Reducing Data Dimension*. (2005) 10-701 S05 April 11 Class Notes. *http://www-2.cs.cmu.edu/ guestrin/Class/10701/slides/dimensionality.pdf*.

[2] J.B. Tenenbaum, V. de Silva, J.C. Langford. *A global geometric framework for nonlinear dimensionality reduction*. (2000) Science, 290:5500:2319-2323.

[3] M. Bernstein, V. de Silva, J.C. Langford, J.B. Tenenbaum. *Graph approximations to geodesics on embedded manifolds*. (2000) Manuscript.

[4] S.T. Roweis, L.K. Saul. *Nonlinear Dimensionality Reduction by Locally Linear Embedding*. (2000) Science, 290:5500:2323-2326.

[5] L.K. Saul, S.T. Roweis. *Think globally, fit locally: unsupervised learning of low dimensional manifolds*. (2003) The Journal of Machine Learning Research, 4, 119-155.

[6] T.M. Mitchell, R. Hutchinson, R.S. Niculescu, F. Pereira, X. Wang, M. Just, S. Newman. *Learning to Decode Cognitive States from Brain Images*. (2004) Machine Learning, 57:1-2:145-175.