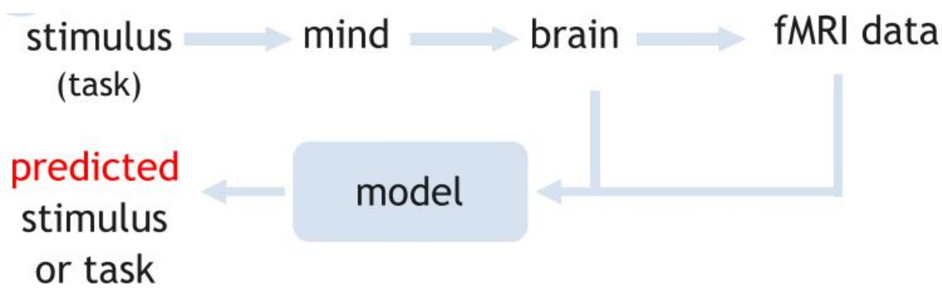
Predicting The Word From Brain Activity

Bhanu Yadav 13198 | Hitesh Menghwani 13310 | Prashant Kumar 13496 | Prince Kumar 13511 | Sawan Kumar 13637

Mentor: Prof. Piyush Rai

Problem Description

In this project, we used data from an experiment where 300 different subjects were given a word from a set of 60 words along with a corresponding line diagram. Each word is associated with 218 human defined attributes. FMRI scan which was recorded of all 300 subjects was used as training data . We needed to learn various models to predict a word (among two candidate words) given fMRI scan of entirely new subject. Our test data consisted of 60 cases.



Through this project we wanted to answer the question does FMRI data of brain carry information about an object, in this case the word and line drawing observed. Can we decode what an individual is thinking about from their brain activity? We use the knowledge that we attained in this course to answer this questions.

Prior Work

The working of human brain has always amazed scientists and they have always tried to decode the functioning of our brain. A lot of work has been and is being done in this field. Some of which inspired us or we took help from for our project are:

- Cox and Savoy (2003) used FMRI data to classify which image category (birds, chairs, etc.) out of 10 the subject was looking at. He used ANOVA feature selection to select voxels out of 27,000 voxels. He used three different classifiers and found that linear support vector machine to be the best one.
- Mitchell et al (2008) predicts FMRI images associated with words for which FMRI data is not available.
- KN Kay et al. (2008) tries to decode which natural image (out of 1000s) the subject is viewing using FMRI data.
- Naselaris et al (2011) tries to predict movie image from the FMRI data obtained while a subjects watched movies.

Data Description

Data provided to us is in the form of a .mat file. We used scipy.io library of sk-learn to extract these. Data has these mappings:
X_train => training set. The training set comprises of fMRI images for 300 subjects. Each fMRI image (given as a 21764 dimensional feature vector of voxel intensities) was captured while the subject was being shown a word along with a "line drawing" representing the word.
Y_train => labels of images in training set
X_test => test set
Y_test => labels for the test set
Word_features_std, Word_features_cnt => values of 218 human defined attributes associated to each of the 60 words in standard and centred forms respectively.

Approach

Initially we tried very simple models like k-nearest neighbours and distance from means to label the test data based on the implicit relations learned using these models. But in this models we were not using a lot of information provided to us like semantic features of the 60 words. However, the maximum accuracy we could get was ~70%.

So, we changed our approach and tried using the 218 human defined features that were provided to us. In this approach we first mapped the FMRI image to semantic features. We used various regression techniques for this task.

A regression model was learnt using the FMRI image training data and semantic features of the word that the image corresponded to. This model was used to predict semantic features of FMRI test image (using the learnt weight matrix). Corresponding to each test image we were given two option words. We used semantic features of those words to find which word's semantic features were closest to the predicted semantic features of the test image. The closer word was the predicted word for the test image data. In short, we mapped FMRI image data to semantic features using regression models and then classified the semantic feature using 1-NN classifier.

Methods Considered and Tried

- **K-Nearest Neighbour**: We applied k-NN (used k = 1 and 3) to classify the given test example into one of the two classes of words. Only those FMRI training images were used for classification corresponding to the two words (one correct and other incorrect) given for a test example.
- **Distance from means:** As in first approach used FMRI training images for classification for which label is one of the given two words using distance from means approach.

For mapping FMRI image data to semantic features we used following methods:

- **Subset Selection:** In this we take all combinations of predictors and fit regression model using those predictors as independent variables. Then we choose the best model out of all the regression models (This can be done by comparing cross-validated prediction error, AIC or BIC). But this cannot be applied to the FMRI dataset as number of predictors is very large.
- **Stepwise selection**: Forward selection begins with a model containing no predictors then we keep on adding predictors till the model improves significantly. Again this method is not feasible for our dataset.

  We then tried using shrinkage/regularization methods which add some regularizer to loss function of standard linear regression model to shrink estimated coefficients towards zero.

- **Ridge Regression**: It uses L-2 regularization to avoid overfitting and shrink the estimated coefficients towards zero but it includes all predictors in the final model. To solve for weight vector we minimize the following loss function:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$
$$\iff \quad \min_{\beta \in \mathbb{R}^p} \underbrace{\|y - x\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

  In our case we learned 218 ridge regression models with independent variables as the 21764 voxels of a FMRI training image and y_i, i = 1,2,....,218 variable as dependent variable. Here y_i represents i^{th} semantic feature of the word that the training image corresponds to. After solving for the weight vector we will get 218 21674-dimensional weight vectors. We can use weight vectors to predict the 218 semantic features for our FMRI test data.

- **LASSO (least absolute shrinkage and selection operator)**
  Lasso is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.
  Lasso regression solves a penalized least squares problem with L1 regularizer:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i^T \beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

$$\iff \quad \min_{\beta \in \mathbb{R}^p} \underbrace{\|y - x\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}} .$$

  Unlike ridge regression it is sparse linear model. The weight vector that is learnt is sparse vector i.e. it will contain many zero entries and a few non zero entries. In our case we will learn 218 lasso regressions and weight vectors. We can use these weight vectors to predict semantic features of test image data.

  It also makes sense to use LASSO model to learn semantic features from our FMRI image data as not all voxels in the brain are activated when a subject is shown word along with line drawing. Only some voxels will be activated and only those voxels will be used to predict semantic features.

- **Elastic Net** : It is a regularized regression method that linearly combines the L1 and L2 penalties of the lasso and ridge methods. To solve for weight vector we minimize the following loss function.

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1$$

  The elastic net method overcomes the limitations of the LASSO (least absolute shrinkage and selection operator) method which uses L1 penalty .Use of this penalty function has several limitations .For example, in the "large D, small N" case (high-dimensional data with few examples), the LASSO selects at most n variables before it saturates. Also if there is a group of highly correlated variables, then the LASSO tends to select one variable from a group and ignore the others. To overcome these limitations, the elastic net adds a quadratic part to the penalty .

- **Partial Least Squares:** PLS is a dimension reduction method. This method constructs new set of features which are linear combination of the original features and then fits the model using the new features as independent variables. In this method new features are learnt in such a way that they not only representation of the original features but are also related to the label i.e. y variable which means that it is a supervised learning method. In our case we implement PLS using plsregress function in Matlab. The number of components were chosen using cross validation.

- **Cross Validation** : It is a model validation technique.The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting, give an insight on how the model will generalize to an independent dataset (i.e., an unknown dataset). LOO(leave one out) was used to obtain Ridge regulariser that would give less test error .

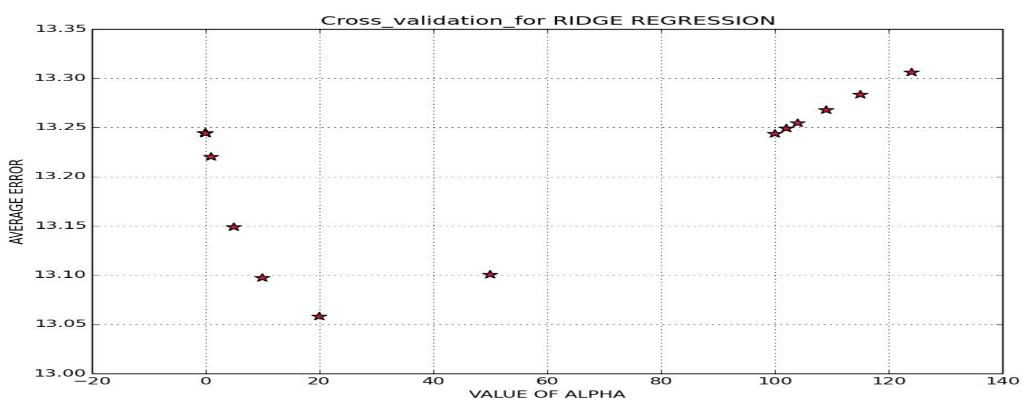
- **Correlation criteria** : This is method to quantify the relation between two vectors . This method was used to get the correlation of each X feature in dataset with each semantic feature Y ,hence these correlation was used to rank the features which was further modeled using Ridge regression .

Tools Used

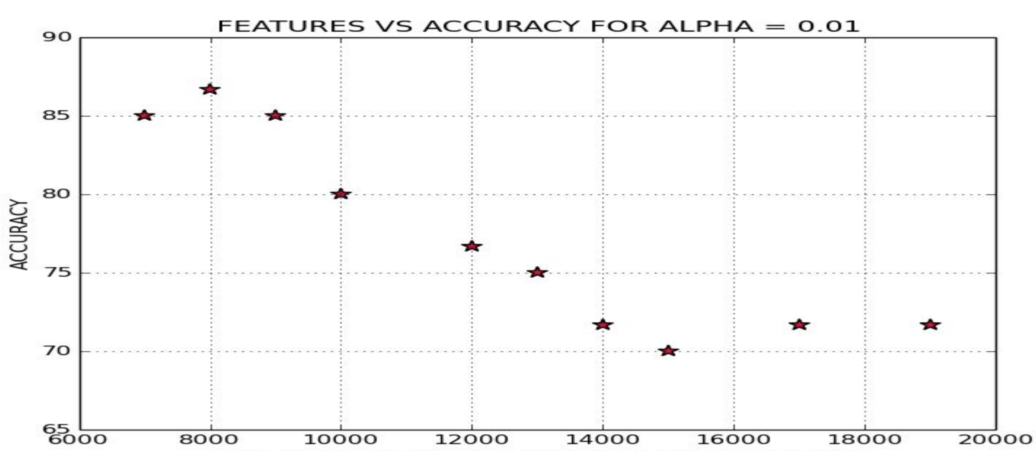
- Python
- Scikit-Learn
- Matlab

Experimental Results

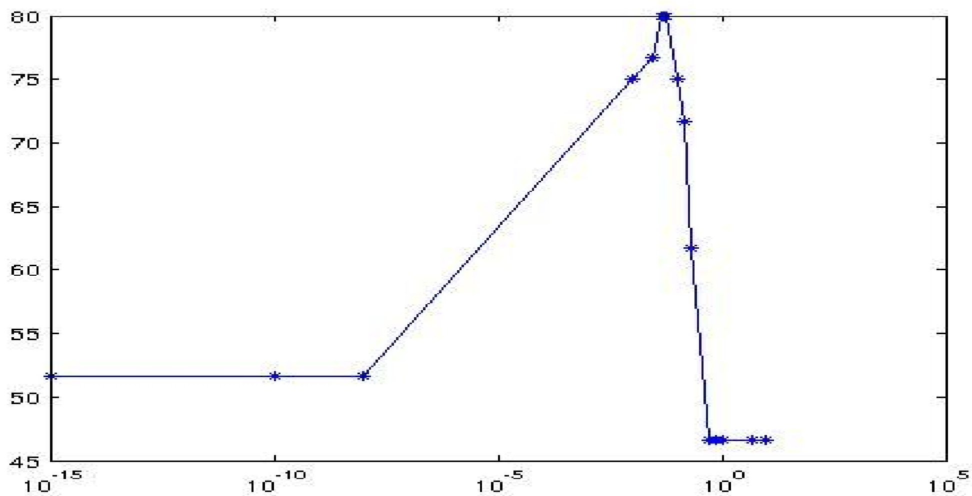
- **1-NN, 3-NN** and distance from means: Though this model used very less information we were able to obtain an accuracy of 63.33 %, 68.33% and 68.33% on the test data respectively .
- **Partial Least Squares and Elastic net:** Although PLS technique is used for predicting labels of high dimensional datasets. However for this dataset this method did not seem to work as we found that it had accuracy. Obtained accuracy of 51.67 % . And surprisingly for Elastic net (for which we expected a higher accuracy) we found accuracy of 68.33 %.
- **Ridge Regression** using alpha obtained from **Cross-Validation**
    - Below graph was obtained when **Cross-Validation technique LOO** (Leave one out) was used to find the best regularisation parameter alpha . We obtained the value of regularization parameter 20 i.e,the value at which error is minimized is at alpha equal to 20. The accuracy obtained on test data using this alpha is 78.33%



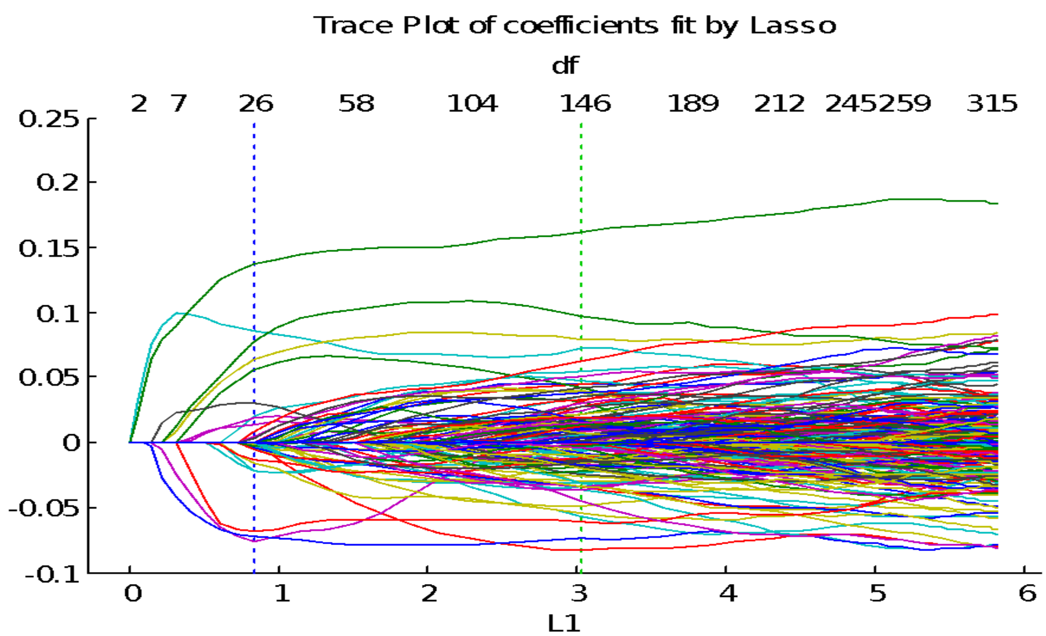
- **Ridge Regression** with feature selection (using Correlation method)
    - Below graph was obtained when **Correlation method** was for feature selection . Data was modeled using Ridge Regression with varying number of features and 0.01 as alpha . Here we got an accuracy of 86% when number of feature selected was 8000 .



- **LASSO:** The graph below was obtained when **LASSO** (alpha(X-axis) vs accuracy(Y-axis)) mode was used to train the data . Maximum accuracy obtained was 80%.



The graph below is **regularization path** of the coefficients in lasso regression. It shows that as the value of L-1 norm constraint or as the value of lambda is increased more and more coefficients shrink to zero.



Things We Learned

- Read Previous research that has been done for fmri data classification and how the authors performed the experiment, methods they applied.
- Learnt how to handle high dimensional data with relatively few observations for regression.
- Different methods that are used like subset selection, stepwise regression, Lasso, etc., when to use these methods.
- Leant about how LASSO objective function is maximized using sub gradients, why it shrinks coefficients of features to zero, different algorithms like least angle regression, stochastic coordinate descent.

References

[1] http://www.cs.cmu.edu/afs/cs/project/theo-73/www/papers/Pujara-MSthesis%20-%205-10-05.pdf -

[2] http://www.cs.cmu.edu/~fmri/papers/Bonarkdarpour_snrthesis.pdf

[3] http://www.cs.cmu.edu/afs/cs/project/theo-73/www/papers/zero-shot-learning.pdf

[4] http://www.cs.cmu.edu/afs/cs/project/theo-73/www/papers/amia2003-final.pdf

[5] Christopher M. Bishop. Neural Networks for Pattern Recognition Oxford University Press, Nov. 23 1995