Predicting Semantic Word Vectors from fMRI Data by Learning Sparse Linear Models

Kruthika Hassan
Department of Applied Mathematics
University of Washington
Seattle, WA 98105
kruthika@uw.edu

Abstract

Cognitive science has made tremendous efforts in understanding how human brain represents conceptual knowledge. A simple analysis of brain images gives valuable insight into neural activation associated with different contexts and categories of words. In this paper, a computational model that reads the mind by predicting the word associated with given fMRI data, is built. Essentially, guessing what's on the subject's mind. The model is trained from a set of observed fMRI data obtained by presenting several dozen concrete nouns to a subject for viewing. The trained model then predicts which word corresponds to a given combination of voxels.

1 Introduction

The question of how the human brain represents and organizes conceptual knowledge has been studied by many scientific communities. Neuroscientists using brain imaging studies have shown that distinct spatial patterns of fMRI activity are associated with viewing pictures of certain semantic categories, including tools, buildings, and animals [1]. To elaborate on what is semantics, it is the study of the meaning of words, constructions, and utterances. Semantics can be divided into two parts, the study of meaning of individual words (or *lexical semantics* and the study of how meanings of individual words are combined into the meaning of sentences (or even larger units). Computational linguists characterize latter part using distributional semantics, that is to analyze the statistics of very large text corpora and construct the distribution of words and phrases with which it commonly co-occurs. This captures the meaning of words and phrases to much extent. Psychologists have also studied word meaning through feature-norming studies [2] in which participants are asked to list the features they associate with various words, revealing a consistent set of core features across individuals and suggesting a possible grouping of features by sensory-motor modalities. This variety of experimental results has led to competing theories of how the brain encodes meanings of words and knowledge of objects, including theories that meanings are encoded in sensory motor cortical areas [3] and theories that they are instead organized by semantic categories such as living and nonliving objects [2].

The theory underlying this computational model is that the neural basis of the semantic representation of concrete nouns is related to the distributional properties of those words in a broadly based corpus of the language. Experimental evidence showing that the best of these models predicts the word represented by a neural activation of the brain. A mapping from an unseen fMRI data into the word it denotes. This also makes imperative that inverse map can be easily obtained once how different contexts and meanings are learnt is understood. The experimental results establish a direct, predictive relationship between the statistics of word co-occurrence in text and the neural activation associated with thinking about word meanings.

An important characteristic of fMRI data is emphasized. The data is represented using voxels, which are a parallel to pixel values in 3-D. Except that the position is not encoded along with the value as

in case of pixels. Voxel representations are high dimensional and have inherent sparsity, which is exploited in this paper.

Rest of the paper is organized as - Section **??** explains the data and the objective of the paper, Section **??** is on theory and algorithm approached, Section **??** enumerates the experimental results, and Section **??** concludes the paper by making a few observations.

2 Problem Setup

Before delving into approach to solution of the problem, a thorough understanding of the dataset that needs to be dealt with and the problem definition needs to be well understood.

2.1 Data

The original data considered in this project is available at `http://www.cs.cmu.edu/afs/cs/project/theo-7/www/science2008/data.html`. It consists of fMRI data collected for several subjects by presenting them with a set of stimulus words and recording the brain activity in response. The words are selected from a vocabulary set consisting of 60 concrete nouns and each word presented 6 times in a random order. Thus, fMRI data is available for 360 trials, corresponding to a single subject.

For each word in the vocabulary set, 12 semantic categories are extracted. But intermediate semantic features are added and extended to 218. Each feature corresponds to a question about the semantic meaning of this word. The value is a score ranging from 1 to 5 provided by a human labeler as his response to the question. For example, feature 2 corresponds to the question: "IS IT A BODY PART". The word "arm" has value 5 for feature 2 whereas the word "ant" has value 1.

At trial i, $word^i$ is shown to the subject. Then the fMRI is recorded as a 21764 dimensional vector X^i. Each dimension corresponds to the activity in a voxel (a cube in the 3-D coordinates of the brain). For a total of 360 trials, the results are put into $X \in \Re^{360\times 21764}$ and $Y \in \Re^{360\times 218}$. $X_{i,j}$ is the signal at jth voxel at trial i, and $Y_{i,j}$ is the jth feature of the word displayed at trial i. X is further standardized to have mean 0 and unit norm for each column and center Y to have mean 0 for each column. The data is finally split into training set (300 trials) and test set (60 trials).

	0	1	2	3	4	5	6	7	8	9	...	21754	21755	21756	21757	
0	0.213785	-0.326802	-0.161990	0.664290	1.168009	1.600899	0.633698	0.398940	0.003676	-0.440629	...	0.631878	0.725080	0.521667	-0.412455	-0.3
1	-0.127519	-0.904420	-0.330786	0.500771	-0.691373	-0.048262	0.005815	0.089271	-0.067566	0.055849	...	-1.055174	-2.433569	-1.577391	-0.547554	-0.2
2	-0.252490	-0.156302	0.271426	-0.320197	0.189128	-0.698400	-0.378035	-0.951903	-0.357715	0.000592	...	-0.033038	-1.027868	0.737419	0.112334	-0.9
3	0.165814	-0.796902	-0.076832	1.709843	1.164824	-0.904719	1.093355	1.165961	0.189203	-0.298003	...	-2.412832	-1.909058	-2.065687	-1.425259	-0.8
4	1.141420	1.276529	0.759836	1.679761	1.118194	0.273017	0.389244	1.039162	1.278746	1.208029	...	1.159715	1.822820	2.094766	1.308743	-0.6

5 rows × 21764 columns

Figure 1: A snapshot of fMRI data

	0	1
0	39.0	7.0
1	17.0	58.0
2	7.0	1.0
3	29.0	47.0
4	26.0	50.0

	0	1	2	3	4	5	6	7	8	9	...	208	209	210	211	212	213	2
0	-0.65	-0.333333	-0.366667	-0.333333	-0.416667	-0.333333	-0.333333	-0.4	1.316667	0.883333	...	0.516667	0.1	-0.066667	0.983333	2.25	1.083333	-0
1	2.35	-0.333333	-0.366667	-0.333333	-0.416667	-0.333333	3.666667	-0.4	-2.683333	-1.116667	...	-1.483333	-1.9	-2.066667	-2.016667	1.25	1.083333	0
2	-0.65	-0.333333	3.633333	-0.333333	-0.416667	-0.333333	-0.333333	-0.4	1.316667	-0.116667	...	1.516667	1.1	0.933333	-0.016667	0.25	0.083333	0
3	-0.65	-0.333333	-0.366667	3.666667	-0.416667	-0.333333	-0.333333	-0.4	1.316667	-1.116667	...	-1.483333	0.1	-2.066667	-1.016667	-0.75	-0.916667	-0
4	0.35	3.666667	-0.366667	-0.333333	-0.416667	-0.333333	-0.333333	-0.4	-2.683333	0.883333	...	0.516667	0.1	1.933333	0.983333	-0.75	-0.916667	0

5 rows × 218 columns

Figure 2: A snapshot of candidate words to be chosen and corresponding semantic vector representation

2.2 Goal

The goal is to learn a set of sparse linear models to predict word semantic features from fMRI signals. Using the predicted features, a binary classifier needs to be constructed such that given two candidate words, it predicts which word the subject was thinking.

To be specific, 218 sparse linear models, each predicting a semantic feature of the output space, $W^i = f(X, Y_i)$, $i = 1, 2, \ldots, 218$. Given a new fMRI data, X^{n+1}, the model will output a 218 dimension vector Y^{n+1}.

3 Approach

3.1 Theory

The data space $\Re^{360\times21764}$ exemplifies the fact that it is a wide data set with features (p) >> number of samples (N). Linear models tend to be the choice for modeling such systems. Since it is an under determined system, we cannot fit these models using standard approaches. A constraining the parameters becomes necessary. While there are various approaches to this like the ridge regression, LASSO, etc, the approach of LASSO is considered here. The advantage of using LASSO over Ridge regression or other methods lies in shrinkage of coefficients towards 0 and also feature selection.

The LASSO solves the following optimization problem:

$$min_{w_0,w} \frac{1}{2n} \sum_{i=1}^{n} (y_i - w_0 - (X^i)^\top w)^2 + \lambda \sum_{j=1}^{p} |w_j| \tag{1}$$

where $X \in \Re^{n\times p}$, $Y \in \Re^n$ and λ is regularization hyperparameter.

The optimization problem is convex, with the regularization part being non smooth and linearly separable in coordinates. These characteristics give an advantage in using the coordinate descent. In coordinate descent, each parameter is optimized separately, holding all the others fixed. Procedure is repeated until the parameters stabilize.

3.2 Algorithm

The algorithm for solving the problem is elucidated as a step by step procedure

1. Read the fMRI train data obtained experimentally during 300 trials into X_{train}.
2. Read the corresponding words represented as a 1×218 semantic feature vector into Y_{train}.
3. Read the fMRI test data obtained for 60 trials into X_{test}.
4. Read the random words and ground truth words into Y_{random} and $Y_{truthwords}$.
5. Fit the LASSO model, $LASSO(X_{train}, Y_{train}, \lambda$ by performing coordinate descent on a grid of values for λ.
6. Stack the predicted words into a 218×60 matrix $Y_predicted$.
7. Calculate the number of correct predictions by finding the L_2 distance of each of the word predictions in $Y_{predicted}$ to each word vector in Y_{random} and $Y_{truthwords}$ as d_1 and d_2 respectively. The correct predictions have $d_1 > d_2$.
8. Choose the best model corresponding to highest accuracy.

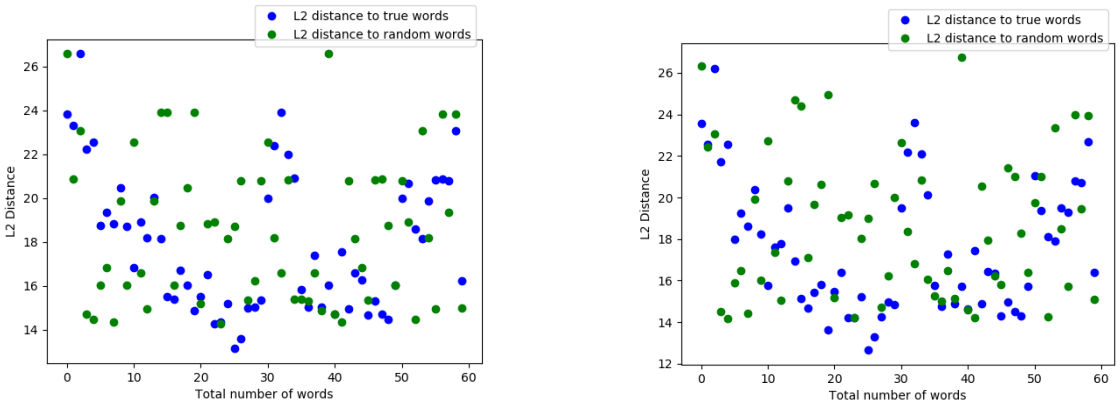
4 Experiment

The results of experimenting with the LASSO for various values of regularization hyperparameter λ is tabulated in Table 1.

The Euclidean distances represented in Figure 3 show that for hyperparameter $\lambda = 1$ and $\lambda = 0.5$, the classification accuracy is not great. We find several random words having lesser Euclidean distance than that of true words.

Table 1: Results for various values of λ

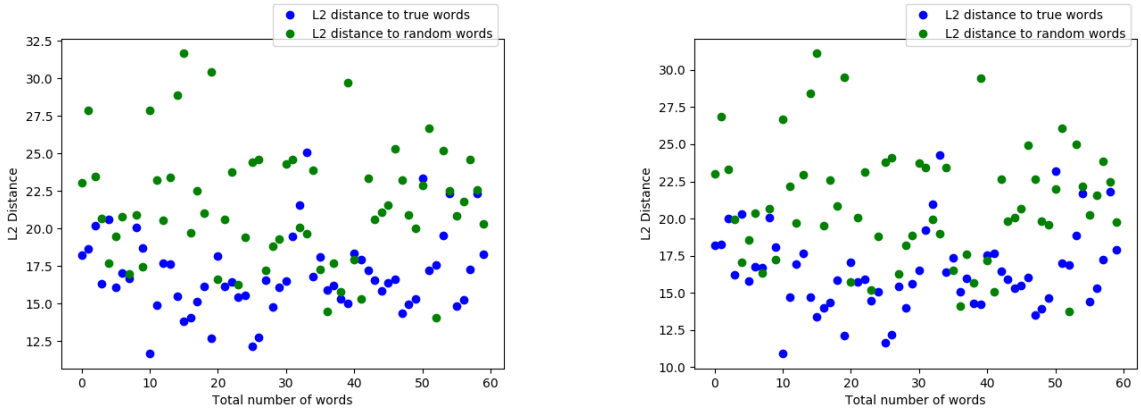
λ	Correctly classified words out of 60	Accuracy (in %)
0.01	49	81.67
0.02	49	81.67
0.04	48	80.00
0.05	48	80.00
0.5	33	55.00
1.0	31	51.67
100	31	51.67
1000	31	51.67



(a) $\lambda = 1$
Accuracy = 51.67%

(b) $\lambda = 0.5$
Accuracy = 55.00%

Figure 3: Plots depicting L2 distances to true and random words



(a) $\lambda = 0.02$
Accuracy = 81.67%

(b) $\lambda = 0.05$
Accuracy = 80.00%

Figure 4: Plots depicting L2 distances to true and random words

In Figure 4, the classification accuracy increases considerably, for $\lambda = 0.02$ and $\lambda = 0.05$. This can be observed from the plot, where an almost discernible demarcation between blue and green dots can be seen.

Figure 5 represents the trend of accuracy with various values of λ.

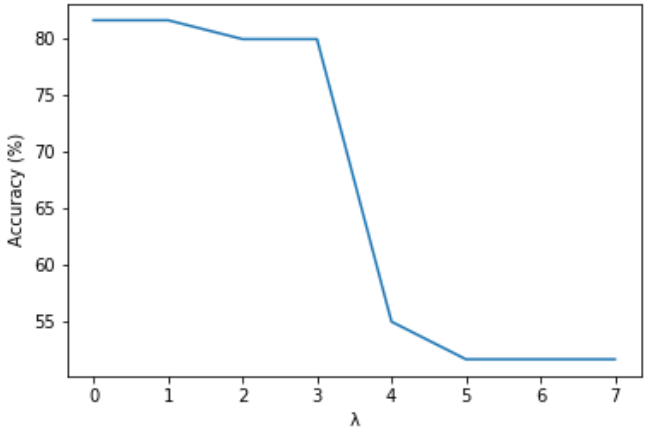


Figure 5: Accuracy vs λ

5 Conclusion

There were a few important observations and conclusions made after the experimentation:

- Linear models are very powerful in fitting data with large number of features and very few samples. This wide data is often encountered in biological studies.
- Solving such under determined systems requires a constraint to be put on the parameters, since there are infinitely many solutions. A LASSO constraint has the advantage of not only low variance but also feature selection. A useful property when the features are huge in number.
- The regularization parameter λ is a hyperparameter that needs tuning to find the best model. Though this is usually done using cross-validation, when the number of samples are small, it is best to tun e the model on a grid of λ values and find the appropriate one.
- It is observed that as $\lambda \to \infty$, that is towards larger values, the regularization term has no effect on the objective, the LASSO is reduced to least squares and the accuracy stalls at a constant value. But as $\lambda \to 0$, tends to smaller values, accuracy increases, though an optimal value needs to be found to avoid overfitting.

Acknowledgments

I extend my gratitude to Dr. J Nathan Kutz for providing the liberty and encouragement to explore each of the assignments and also project.

References

[1] Mitchell, Tom M & Shinkareva & Svetlana V & Carlson et al.., (2008) Predicting human brain activity associated with the meanings of nouns, *Science, American Association for the Advancement of Science*

[2] Cree, George S & McRae, Ken (2003) *Journal of Experimental Psychology: General*, US: American Psychological Association.

[3] Martin, Alex & Chao, Linda L (2001) Semantic memory and the brain: structure and processes. *Current opinion in neurobiology, Elsevier*.