Classification of 'Honest' and 'Deceitful' Memory in an fMRI Paradigm
CS 229 Final Project
Tyler Boyd-Meredith

Introduction

Background and Motivation

In the past decade, it has become popular to use classification techniques from machine learning in the interpretation of EEG and fMRI data. This is often called 'multivoxel pattern analysis' (MVPA) by the neuroscientists who use it, and has been applied with varying degrees of success in a wide range of areas. The appeal of the technique is twofold. First, it can lend insight to those who are interested in how mental states arise from distributed patterns of activity across the brain. Second, it provides the opportunity to attempt 'mind-reading'. While exciting, the second appeal has opened the door to many worrisome applications of fMRI. In California, at least two companies, 'No Lie fMRI, Inc.' and 'Cephos ,Inc.' have cropped claiming to provide lie detector tests using neuroimages. In India, over 1000 court cases have in some way relied on evidence from a test that can supposedly detect the presence or absence of experiential memory for information that only the perpetrator could know.

Dr. Melina Uncapher is leading a project to illuminate the issues involved in memory state decoding as applied to lie/experiential detection, as well as the basic science behind memory encoding and retrieval. Rissman (2010) has already demonstrated that one could train a classifier to recognize whether a subject was subjectively experiencing an image of a face as old or new after having attempted to learn a large set of face images. Dr. Uncapher's project attempted to (1) replicate these findings, and (2) investigate whether or not people might be able to volitionally manipulate their brain responses when motivated to conceal or feign recognition of a stimulus. In other words, can the subjects beat the classifier using 'countermeasures'?

Machine Learning Questions

I became involved in this project during the data collection phase and was interested in finding out what else could be done with the data. I asked the following questions:

1. What is the best algorithm to use for binary classification of this data? Penalized logistic regression (pLR) is used frequently in this literature, and I wondered if we might get better performance using a different penalty or using support vector machines.
2. Would we be able to correctly classify subjects memory states in the countermeasures condition? How might different training and testing schemes affect out performance.
3. How might expanding the scope of the problem to multiclass classification using softmax regression allow us to mitigate the problem of countermeasure deployment?

Data Collection and Featurization

Study Description

- Encoding (Day 1): Participants (N = 24) were shown a set of 200 faces (each subject saw a different image set). Images came from a database, were standardized, gray scaled, with all white background. Each face appeared for 2s and participants saw each face twice.
- Retrieval (Day 2): Participants were scanned while making recognition judgments about the 200 faces seen at encoding (OLD) and 200 novel faces (NEW). Each subject did 200 explicit (EX) trials and 200 countermeasures (CM) trials. In each explicit trail, the subject

was asked to respond with their honest recognition judgment (if they remembered the face, they responded 'old', and if they did not remember the face, they responded 'new'; responses were recorded via button box). In each countermeasures trial, the subject was asked to respond with the opposite answer (if they remembered the face, they responded 'new' and if they did not remember the face, they responded 'old'). In addition, subjects were also instructed to attempt to volitionally change their patterns of brain activity when they saw the stimulus. In particular, when faced with previously encountered images, they were instructed to focus on unremembered, or previously unnoticed aspects of the image, for example, the exposure of the image, the shadows, etc. On the other hand, when faced with novel images, they were instructed to ascribe memories to that image, for example, they could make up a story about the face belonging to their neighborhood ice cream man, and recall the last time they ate from the ice cream truck, etc. The idea was to make the pattern associated with novelty look more like the pattern associated with recognition and vice versa.

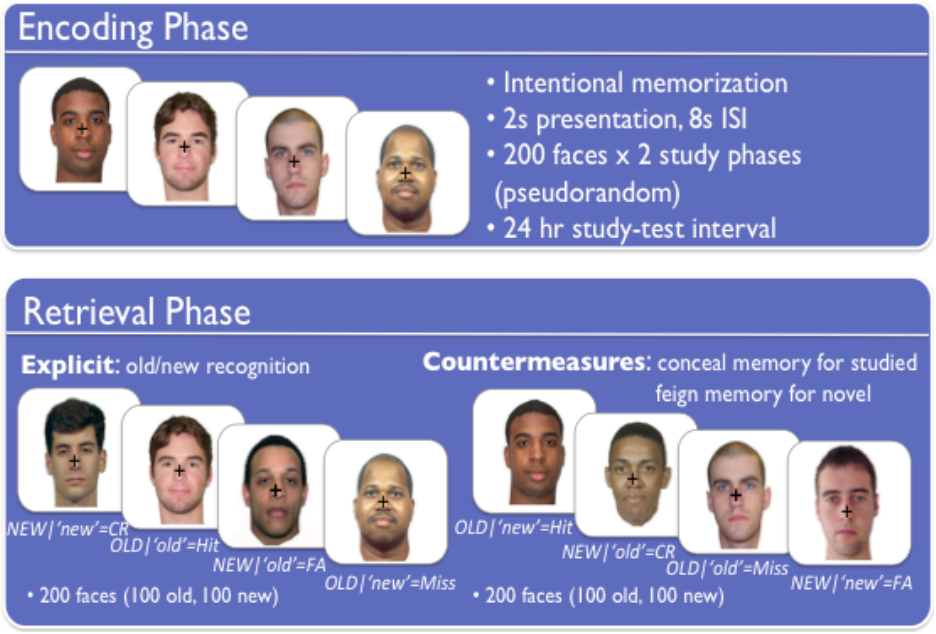

Figure 1. At encoding each subject (N=24) saw all 200 faces twice for 2 seconds, with 8 seconds in between. At retrieval, subjects saw the same 200 faces encountered at encoding, as well as 200 novel faces. For the first 200 trials, they were asked to respond honestly (EX) and for the final 200 trials, the subjects were asked to respond falsely (CM) and employ countermeasures so as to make their brain look like it was remembered when it wasn't and vice versa.

Previously validated procedures suited to collect fMRI data for MVPA were used to obtain the neuroimages that were used for classification. In particular, the interscan interval was set to 8 seconds and the activation intensity values given to the classifier were the averaged values from 3, 4, and 5 seconds after stimulus presentation. These timepoints gave us the best classification accuracy (results not presented), which is expected given the nature of the blood oxygen level-dependent (BOLD) signal used for fMRI.

Additionally, though the activity of over 40000 voxels was recorded, a 23000 voxel mask was applied to exclude the cerebellum, motor, premotor, and somatosensory cortices. The reason for this is to avoid giving a degenerate solution to the classifier. Because subjects are responding by button box, any brain activity correlated to their responses would directly indicate the correct answer for the classifier.

The data was labeled using the veridical status of the images (whether the image had or had not appeared in training), the subjective status of the images (whether the image had been

recognized correctly (hit), recognized incorrectly (FA), forgotten (miss), or correctly identified as new (CR)). Note: responses obtained during countermeasures trials were flipped, so a response 'new' when the stimulus was OLD, was marked as a hit, etc (see Fig 1).

Results

Penalized Regression on Explicit Trials and Countermeasures

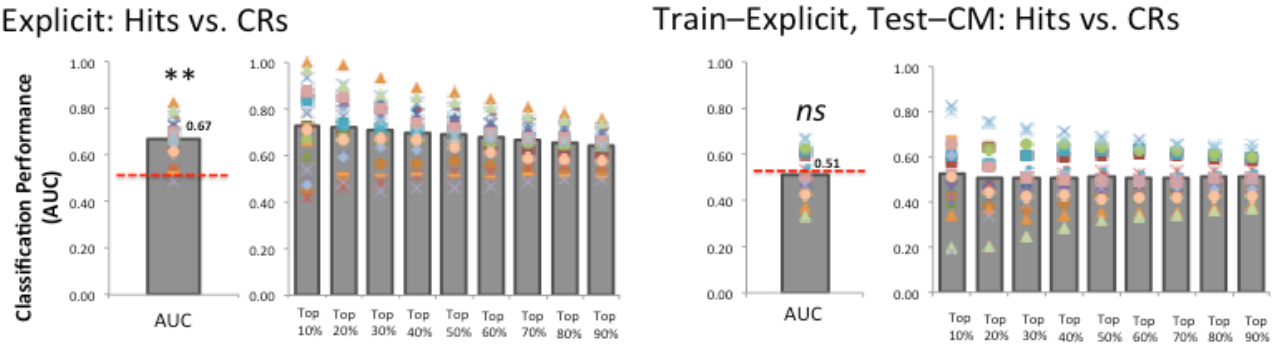

Figure 2. Area under the ROC curve (AUC) values for discriminating hits from correct rejections when training and testing on explicit trials (0.67) and when training on explicit trials and testing on countermeasures (0.51). Gray bars represent the mean value across subjects, while points represent the mean for each subject (N=24).

Classification was performed with heavy reliance on the Princeton MVPA Toolbox, a library, which provides much of the skeleton for logistic, softmax, and svm classification algorithms (among others), as well as some wrapper methods, which help featurize the data, written for use in our lab by Dr. Rissman.

The first two classifications were both trained to discriminate hits from correct rejections using explicit trials as training data. The first tested on explicit trials and the second tested on countermeasures trials. The algorithm used was penalized logistic regression (pLR) with penalty = 10. Performance for training and testing on each subject was quantified using area under the ROC curve (AUC), which was obtained using the average of 10 4-round leave k-out cross-validation schemes (k varied depending on the test).

We had better than chance performance when training and testing on explicit trials. Interestingly, we had chance performance (AUC = 0.51) when testing on countermeasures trials. This strongly suggests that subjects were able to *volitionally* manipulate their retrieval strategies.

Though we had performance far above change in distinguishing between correct recognition (hits) and correct rejection of novel faces (CRs) (AUC = 0.67), we had hoped for even better performance. I offer the interpretation that because so many of our trials were countermeasure trials, which we could not use to train on for this purpose, we should expect worse classification data than experiments that had achieved higher values (Rissman 2010), since they had only explicit trials to train on. Regardless, the first task was to validate the pLR scheme.

Validating Penalized Logistic Regression

To discover whether or not performance could be improved with another algorithm, PLR was tested with penalty = 5, 10, and 15, and SVMs were tested with linear kernels and polynomial kernels of degree 2 and 3 (Figure 3). These tests were done using only 5 of the subjects for the sake of time. Of the penalties tried, the best pLR performance as associated with penalty = 10. Of

the SVMs used, the degree 2 polynomial kernel performed the best, however it did not outperform pLR.

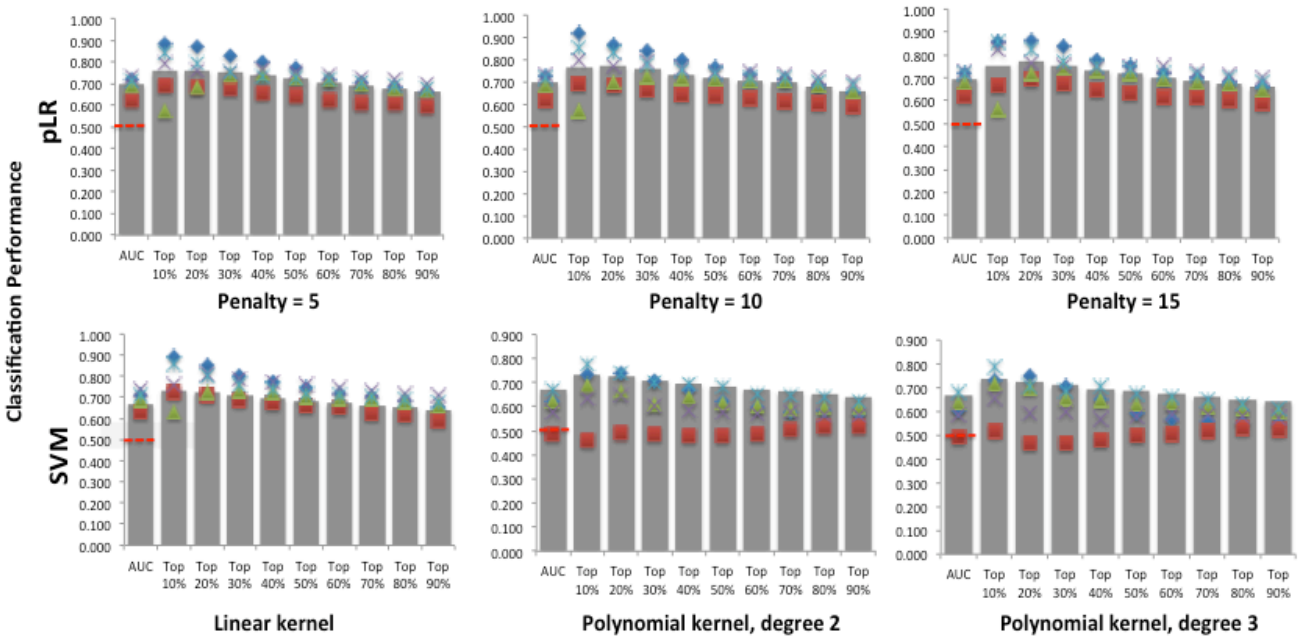

Figure 3. PLR and SVMs were tested on the data with varying parameters to see if any algorithm was better than the others. Penalties used were: 5, 10, 15; SVM kernels used were: linear, polynomial degree 2, polynomial degree 3.

New Training and Testing Schemes

Varying the parameters of the previous algorithms did not improve performance on the explicit trials, so pLR was used for all subsequent binary classifications. Subsequent binary classifications attempted to mitigate the problem of classification in the face of countermeasures, by changing the training and testing schemes (Figure 4). Again, these tests were done using the data from only 5 because of compute time concerns.

The first training and testing scheme was very successful, but flawed. The first attempted was to train on all trials and discriminate between explicit and countermeasures, irrespective or hits, etc (AUC = 0.82). This high performance may owe to the fact that explicit trials and countermeasures trials contained different proportions of hits and CRs (this stems from the fact that the countermeasures condition is difficult and detracts some attention away from the recognition task).

The next training scheme trained and tested on all trials and discriminated between hits and CRs. In this case, we did not have to worry about the problem of proportions, and performance was better than in the train explicit test explicit condition (AUC = 0.71; likely owing to the larger amount of data—trained on 100 EX trials, and 100 CM trials, and tested on the other 100 EX trials and 100 CM trials—though it is a surprise that more data beat the fact that some of the data was countermeasures trials).

The final training, testing scheme tried, attempted to discriminate between hits and CRs in the countermeasures trials (testing on 100 CM runs, after training on 50 EX runs and 50 CM runs).

This performed quite well (AUC = 0.68) compared with the attempt to train on explicit trials and test on countermeasure. A good interpretation of this might be that training on explicit trials overfits the weights, making them inappropriate for the countermeasures conditions. By training on a combination of conditions, it may be the case that the weights have captured distributed patterns of activity that occur regardless of whether or not countermeasures are employed. We may find it very informative to understand which areas of the brain are strongly weighted by the algorithm when trained on the mix of explicit and countermeasures trials. These will be overlaid on the image of the average brain of our participants and examined at by more experienced neuroscientists from my lab in the coming weeks. For this finding to be relevant in the applied context, it will also be necessary to determine if we can get good performance when training on groups and testing on individuals, but preliminary results from Rissman (unpublished), suggest that this is the case.

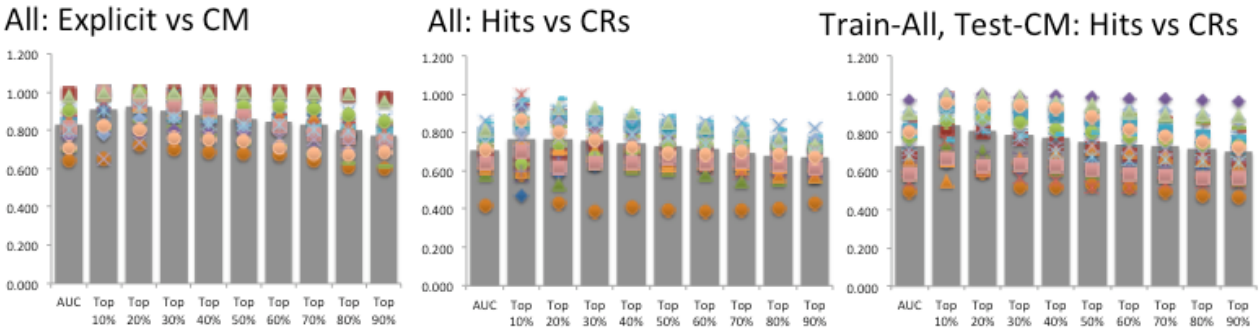

Figure 4. AUC and overall accuracy by classification confidence for (1) training and testing on all trials, discriminating between explicit trials, and countermeasures trials, (2) discriminating between all hits and CRs (regardless of EX or CM), and (3) training on all trials and testing on CM, discriminating between hits and CRs.

NB: Softmax Regression I also implemented softmax regression. I don't have space to present the findings in detail, however I will note that the classifier performed far above chance, discriminating best between EX hits, EX CRs, CM hits, and CM CRs (0.47, 0.41, 0.59, 0.57).

Acknowledgements

This project would not have been successful without the hard work of Tiffany Chow (formerly of the Wagner Lab, now at UCLA), and Dr. Melina Uncapher (of the Wagner Lab), who collected the data and brought me onto the project as they were finishing this process.

The project would not have been possible without the Princeton MVPA Toolbox, which provides a large library of MATLAB scripts tailored to analyzing brain data.

Thanks also to: Dr. Jesse Rissman, Dr. Anthony Wagner, and the MacArthur Foundation Research Network on Law and Neuroscience, which funded the study.

References

Rissman J, Greely HT, Wagner AD. Detecting Individual Memories Through Neural Decoding of Memory States and Past Events. Proc Natl Acad Sci. 2010; 107(21):9849-54