CS 7616 - Fall 2014 - Pattern Recognition - Problem Set 1

Professor Aaron Bobick
Daniel Castro - 902641790

Part 1: Alcoholism Dataset

For the alcoholism dataset, Bayesian Estimation and Maximum Likelihood had similar performance. The results were not particularly accurate so I examined the dataset manually, and concluded that it was in my interest to simply remove the last data column which contained the average number of drinks the participant had per day. I modified the data and achieved negligibly better results, which simply demonstrates that the column neither helped nor hurt the classification. I was more heavily inclined to rely on scientific values rather than ones that were reported by the participant (Although I could not find a precise description of where the 'number of drinks' was obtained, I assume they asked their participants). Running MLE on a dataset of 5 samples often resulted in a singular matrix (could not obtain the inverse) so we ran the training set on seven samples instead. The following were my results:

	Maximum Likelihood Estimate	Bayesian Estimation
Training Set (Averaged) (7)	0.4952	0.5205
Training Set (Averaged) (50)	**0.5977**	**0.5909**
10-fold Cross Validation	0.5694	0.5766

Description of the results:

The results were relatively poor as both methods were only able to classify half of the dataset. The 10-fold cross validation is an accurate depiction of the average results, as the choice of training sets was randomized, which made the accuracy results relatively volatile. We averaged our training set samples over 30 iterations to account for this volatility. I believe that part of the issue is the classification problem seems somewhat arbitrary. The classification of a data point as an alcoholic or not based on the amount of liver damage and/or the number of drinks is not an accurate representation of the classification. Someone who has recovered from being an alcoholic could still suffer from severe liver damage which begs the question of this classification, but I am not well versed enough in the medical field to assess whether or not there are other parameters that may facilitate this classification problem.

Use of code:

I have taken a concept from a StackOverflow article in which someone explains how to obtain the average of the rows of a matrix, which I used to obtain the average of my data. The line of code is as follows:

Adapted from http://bit.ly/1iRa4Xj

```
self._avg_0 = np.matrix([sum(val)/len(val) for val in itertools.izip(*(self._full_training_data))]).T
```

Other Details:

It is also the case that Bayesian Estimation was unable to perform significantly better in this case because we did not have strong priors, which is the whole purpose of using this method. In this case, I only used the average over my entire training dataset (not the entire dataset) as I felt the results I would obtain from using the entire dataset and then testing on that dataset would lack validity. Lastly, for all of my experiments I made sure my 10-fold cross validation contained an equal distribution of each class per fold.

Part 2: Data Banknote Authentication

For the data banknote authentication dataset, Bayesian Estimation performed rather poorly in comparison to Maximum Likelihood. These results were not consistent in comparison with my classmates which indicated to me that I could have committed an error in computation. Given the performance of MLE, I decided that the data did not require any modification. This dataset was obtained from the provided list of datasets, and can also be accessed in the UCI Machine Learning Repository.

	Maximum Likelihood Estimate	Bayesian Estimation
Training Set (Averaged) (5)	0.7643	0.5805
Training Set (Averaged) (50)	0.9857	**0.6584**
10-fold Cross Validation	**0.9873**	0.6551

Description of the Results:

These results indicate a much better performance by MLE. We see an expected increment in accuracy as we increase our training set size, and cross validation performs negligibly better for a much larger training dataset. The training samples were also averaged over 30 iterations in order to account for the volatility of picking a limited number of samples. In this case it is difficult to say what a good prior for the features of a wavelet of an image would be. We can account for the poor performance of Bayesian Estimation due to the lack of a representative prior. As noted previously, we took our priors from the entire training dataset because I felt that using the dataset we are testing on to obtain our priors would void the validity of my results.

Use of code:

As stated in Part 1, we adapted a line of code from a StackOverflow article, see above.

Part 3: Skin Segmentation Dataset

For the skin segmentation dataset, we were provided with 4 attributes, 250,000 data points, and two classifications. Due to the massiveness of the dataset, we took a subset of this dataset (included) and processed 5,000 data points per class for a total of 10,000 instances. The attributes from the dataset are obtained by randomly sampling RGB values which is prone to noise. In retrospect I would perform the data collection differently by obtaining an average color of patches to combat noise although it seemed like an incredibly difficult problem to begin with. The dataset does not state how they obtained non-skin patches, but with more time, I believe the average of super pixels in an image may be a better approach to achieve better accuracy.

	Maximum Likelihood Estimate	Bayesian Estimation
Training Set (Averaged) (5)	0.8616	**0.5561**
Training Set (Averaged) (50)	**0.9850**	0.4878
10-fold Cross Validation	0.8333	0.5310

Description of Results:

In this dataset, MLE once again outperforms Bayesian Estimation. However, it is interesting to note that even after averaging MLE over 30 iterations for 50 random training samples, I obtained an accuracy of 98.5%. Given the lack of precision of these results even with averaging a random set of training samples, it leads me to believe that this is indicative of how the data was collected. The pixels were picked at random which makes the results be volatile. I believe that given the range of skin color, it is once again difficult to have a prior when your only feature points are three color channels. This makes the development of a prior for Bayesian Estimation affect its performance in comparison to the more robust MLE.

Final Remarks:

I believe in the computation of my priors I made a mistake that lead my results in Bayesian estimation to perform significantly worse than expected. I compared these with peers and it was often the case that they would perform relatively similar, but I was unable to pinpoint the potential bug.

Works Cited:

1. Bache, K. & Lichman, M. (2014). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].

Irvine, CA: University of California, School of Information and Computer Science.

2. Hastie, T, Tibshirani, R. Friedman, J. (2013). The Elements of Statistical Learning. Springer. Print.
3. Murphy, K. (2012). Machine Learning. Massachusetts Institute of Technology. Print
4. Duda, R. Hart, P. Stork, D (2001). Pattern Classification. 2nd Edition. John Wiley & Sons. Print.