Week 5 Homework

Samuel Cuthbertson SAMUEL.CUTHBERTSON@COLORADO.EDU

1. Answers to Part A

1.1. Smaller hypothesis spaces tend to have higher Rademacher complexity than larger ones.

False. Since the *maximum* Rademcaher complexity grows proportionally to the size of the hypothesis space, larger hypothesis spaces tend to have larger Rademacher complexity.

1.2. The VC-dimension is the maximum number of points that can be shattered in an infinite number of ways.

True. For example, the VC-Dimension of a line shattering points on a plane is 3, since all combinations of 3 points on a plane can be shattered an infinite number of ways by a line.

1.3. VMs, logistic regression, and perceptrons are all examples of "families of functions."

True, since linear classifiers were defined as synonymous with families of functions in lecture.

1.4. SVMs, Nave Bayes, and logistic regression all find a hyperplane to separate data.

True, since they are all linear classifiers and that is what linear classifiers do. Naive Bayes is a bit of an oddball, but this stack overflow post does a great job explaining how it is in fact a linear classifier in a specific feature space.

1.5. Which has higher entropy? A rare word or a common word?

A rare word, since it takes more bits to store when using a something like morse code as a mechanism for compression.

1.6. What is a Rademacher variable, and what function does it serve in terms of determining Rademacher complexity?

A Rademacher variable is a σ where

$$\sigma = \begin{cases} 1 & with\ probability\ .5 \\ -1 & with\ probability\ .5 \end{cases}$$

Rademacher variables are used to see how a completely random classifier perform, and is useful in determining the performance of a given classifier.

2. Vowpal Wabbit Work

I split the data into a training set of 1000 and a test set of 197 examples for this, and used the attached scripts `vwtrain.sh` and `vwtest.bash` Using a hinge loss function, VW only correctly classified 107 of the test examples, or only 54%. This is in direct contrast with when using either `--ksvm` or a logistic loss function, where VW correctly classified 197 out of 197 examples. When using l2 regularization with a lambda of 1, VW also correctly classified 100% of the examples.

I'm not familiar enough with loss functions, l2 regularization or VW to understand what this tells us about the data, but there is clearly something either wrong with hinge or incorrect in my testing methodology.

3. Two Dataset Projects

3.1. First Project - Ship Nationality

A project that interests me involves the Ocean Ship Logbook dataset, and is classifying the nationality of ships based off of other features in the dataset (Log Notes, Locations, Dates, etc...). This is an interesting project because of the dependent nature of features and classes, and is very attractive to me.

3.2. Second Project - Climate Data

The other project that really intrigues me is analysis on a new coder survey. I'd be curious to see the correlations from various sources to getting a job in industry, and from the impact different backgrounds have on what methods people used to learn. This is less of a classification project and more of a data analysis project, but would be interesting nonetheless.