Weekly Homework 5

Alex Ring

September 23, 2016

Part A

1. T/F Smaller hypothesis spaces tend to have higher Rademacher complexity than larger ones. Explain.

False. The more features you have, IE a larger hypothesis space, the higher your rademacher complexity gets. From the slides in class we can see.

$|H| = 1 : h(x_i)E_\sigma[1/m \sum_i^m \sigma_i] = 0$ and $|H| = 2^m : m/m = 1$

Therefore, rademacher complexity is larger for more complicated hypothesis space.

2. T/F The VC-dimension is the maximum number of points that can be shattered in an infinite number of ways. Explain.

True. The V/C dimension is a way of measuring how well something is at shattering the data. The VC space of shattering data on the interval [a,b] can be classified as ≥ 2 because with two points you can completely shatter the data every time. But with 3 points, no set of three points can be shattered. So the VC dimension of intervals is 2 or the largest amount of data that can be shattered on an interval in infinite ways. Similarly the VC of the sin function is infinite because there is no amount of data points that cannot be shattered using it.

3. T/F SVMs, logistic regression, and perceptrons are all examples of "families of functions." Explain.

True, all of these classifiers are doing essentially the same thing. They are taking some sort of input, training a model, and then classifying yet unseen data based on probabilities. You could refer to these three functions at the "family" of probability classification. They are a family of functions because they are doing very similar things to build models and classify data.

4. T/F SVMs, Naive Bayes, and logistic regression all find a hyperplane to separate data. Explain.

False. Naive Bayes and Logistic Regression use a linear model to train a feature set and then use probabilities based on this feature set to guess which class the test data lies in. SVMs on the other hand try to separate the data with a hyperplane and then classify based on this hyperplane. It is fundamentally different than how Logistic regression and Naive Bayes attempts to classify data. Bayes and Regression use a linear model based on probabilities, while SVMs try to fit with a hyperplane.

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

In a decision tree a rare word is going to have higher entropy. Assuming a large training set, a common word would most likely be able to be classified perfectly into a decision tree. While a rarer word would be much harder to classify and have a much larger entropy. A common word would be expected to provide less information to the system and therefore have lower entropy. While a rare would would provide more information to the system and therefore have higher entropy.

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

The Rademacher variable is a random variable that has equal probability of being 1 or -1 with probability of 50% respectively. It helps in the rademacher function to determine how good a classifier is at classifying completely random data. With this method you can determine if your classifier is just getting lucky. You may be right for stupid reasons.

Part B

Using Vowpal Wabbit, using the data provided for Programming Assignment 2, run the experiment with at least two different loss functions (or neural networks or SVMs) and compare the results. How does using L2 regularization affect the results?

I have included the Vowpal Wabbit Data I used for these tests in the zipfile. I used 80 percent of the to train the model, and 20 percent of the data to test the model. I then measured the accuracy and average loss using the squared loss function and the logistic loss function, with and without L2 regularization.

Overall both functions performed equally without l2 regularization, with an average loss of 0 and an accuracy of 100%. With L2 turned on we see the average loss increase for both functions, with the increase being larger with the squared loss function. We also see that

Table 1: Average Loss and Percent

	Logistic No L2	Logistic With L2	Squared No L2	Squared With L2
Average Loss	0.000000	0.410256	0.000000	0.625000
Accuracy	100%	58.97%	100%	37.5%

the accuracy plummets for both functions, with the logistic L2 outperforming the squared L2. In my testing using L2 with this causes the functions to lose more and over fit.

Part C

Go to https://www.kaggle.com/datasets and browse the data sets.

Describe at least two projects you would be interesting in tackling and why. (The need not be from Kaggle data sets.)

One of the projects on Kaggle is called Death in the United States. It describes a bunch of different death records and some of the data surrounding them. This includes things like age, residential status, marital status, education level, how they died, place of injury, race, ect. For a ton of different deaths in the United States. I think it would be really interesting to look at this data set and train a model to measure the likelihood someone is to die based on certain things in their life (disease, smoker or not, where they live, education level, age, ect). It would be interesting to try to fit a model to the data and then predict the likelyhood of death based on a persons live as a feature set. It is pretty morbid but I think it would be extremely interesting. https://www.kaggle.com/cdc/mortality

Another project I would be interested in the voice gender dataset. This project collects a bunch of voice wave samples and classifies them by gender. This includes various data related to different voice samples and which gender they belong to. It looks like the researchers for this dataset have already attempted a few different models including Logistic Regression, CART, Random Forest, SVM and XGBoost. They have determined that CART is the best kind of model. I would be interested in seeing if we could either improve on their model, or find data that can trick the model. It would be interesting to see how this model would handle data from children, who's voices have not really been developed yet. https://www.kaggle.com/primaryobjects/voicegender