Assignment 1: Introduction to Data Mining

James Quacinella

1. Briefly explain, in your own words, the difference between supervised and unsupervised learning. Give examples to support your explanation. (6 points)

The difference between supervised learning and unsupervised learning is whether or not the data set being used includes the 'correct' value for the response variable. For example, if we want to make product or music recommendations for users, this would be unsupervised: the whole point is that we do not know up front what the users like, and we need to figure that out from data. In this case, clustering users into groups can help find products user may like based on their cluster membership. On the other hand, if we wanted to know what properties of users tend to predict whether or not a user subscribes to a premium service, that would be supervised, since for every user, we know if their are a subscriber or not.

2. Describe a classification problem that would be of interest to you. As part of your description, describe the data you would need in order to solve your problem. Be sure to discuss the types of variables included and any issues you would anticipate with quality of the data. (6 points)

One classification problem that I would be interested in (which I mentioned in the discussion post) would be categorizing short snippets of text, like tweets or sentences. The categories may or may not be pre-determined (i.e. try to cluster them and see if the clusters group them by some theme, or try to match a tweet to a pre-determined set of clusters like politics, products, etc).

The data needed would be a source of these snippets of text. This data may be post-processed to generate other attributes, like which words in the sentence are more important, or to eliminate known useless 'stop words'.

Using something like Twitter can be problematic, since there are many typos in tweets, as well as shorthand, emoticons, and sarcasm. This means extra filtering in post-processing step.

Using a different data set, like new headlines, would be better due to the higher quality of input data. Also, using news headlines might allow us to know ahead of time what category the headline fits into, based on the section of the publication it came from (i.e. an article in the politics section of the NYT). This would lend itself more towards supervised learning.

3. Describe an estimation problem that would be of interest to you. As part of your description, describe the data you would need in order to solve your problem. Be sure to discuss the types of variables included and any issues you would anticipate with quality of the data. (6 points)

An estimation problem everyone is trying to work on is predicting / estimating what the value of some financial product, like a stock. Designing a model to help estimate what future worth a stock will have is difficult, as there is no good underlying theory as to why stocks move the way they do, and human actions (individually or collectively) that affect price are hard to predict and measure.

Issues with the data revolve around the fact that there are a lot of variables to worry about, including the previous worth of the stock over time, volatility in the market, recent news about the company, recent news about that company's industry, etc. Some of this data is easy to obtain and should be of good quality. Other parts of the data themselves are interesting to model (like how news can affect price movements). Some companies have tried to analyze social media feeds for hints about the general economic mood, or the feelings about a certain company, to help predict price movements in the market. While interesting, this data is fuzzy at best.

4. On page 52 of the book An Introduction to Statistical Learning, there are three examples of problems in data mining. (See exercise 2.) For each, explain whether it is a classification or regression problem and indicate whether it is a problem of prediction or inference. (Do not worry about n and p just yet.) (6 points)

The data set revolves around US businesses, and the relationship between CEO salary (a quantitative variable) and properties like profit, number of employees, industry (a qualitative variable). This would be a regression problem, since the response variable is quantitative. This is a problem of inference, as we would like to know which variables affect the CEO salary. It would be unlikely we can come up with a way to predict the exact value of the salary when one of the variables is qualitative.

a. In this case, we want to predict whether or not a project will be a success or failure, so this is not an inference problem. This is also a classification problem, since the response variable is qualitative (i.e., we need to classify an input instance into one of two groups, success and failure).
b. Since we are trying to predict the % change in the US dollar as compared to the world stock market, this is a prediction problem. The response variable is quantitative, so this is a regression problem.

5. Some problems are best solved by a learning approach, while others are better handled by a design approach. Describe a problem of interest to you that can be handled either by learning or design. Which approach would you more likely choose? Why? (6 points)

In a previous class, we learned how to create a math model that can help predict vehicular stopping distance. We can use both methods to solve this problem. In that course we went the design way: using physical intuition, we came up with what a model would look like. However, we could have used a learning approach. This would have entailed driving cars, applying the brake at a certain time and seeing the distance the car travels until it stops. Using the tallied data, we could use a supervised learning method to learn a function f() that would predict the stopping distance. I would lean towards the design / modeling approach: since we know the underlying physics, it seems more natural to work this way. Also, coming up with an apparatus to collect this data would be problematic and costly.

6. We will be using R as our primary tool for the course, but we will also be exploring other tools. In particular, you will be required to do some work in two other environments, chosen from MatLab, Python , and Microsoft Azure Machine Learning. (5 points)

Python, since I am very familiar with it, and Microsoft Azure Machine Learning, since I am interested in cloud ML solutions (even if its Microsoft).

7. Complete the introductory survey in the discussion forums! (5 points)

Done