CS6923 Machine Learning, Spring 2015
Prof. Hellerstein, NYU Polytechnic School of Engineering

## Homework 1

**Due Mon. Feb. 23 at the start of class.**

Note: If you do not already have Matlab installed on your computer, you should get it installed by going to laptop support in Rogers Hall. On NYU classes under Resources, I have posted links to some helpful Matlab tutorials. The Matlab programming for this homework is easy. You will use more Matlab in later homeworks.

1. Let $V$ be the set of points $x \in \mathbb{R}$ satisfying the inequality $x \geq 0.7$.

   0.3 Let $I$ be the closed interval $[0, 1]$, consisting of all real numbers between 0 and 10.

   (a) Suppose we are given 500 points $x \in \mathbb{R}$ that are drawn uniformly and independently at random from the real-valued points in the interval $[0, 1]$.

   Of the 500 points generated, what percent do we expect to be in set $V$?

   (b) Write a Matlab program that repeats the following experiment 10 times: Generate 500 points $x \in \mathbb{R}$ uniformly at random in the interval $[0, 1]$ using the Matlab rand() fucntion, and then calculate the percentage that are in $V$.

   Your program should then do the following:

   i. Produce a bar graph, with 10 bars, one for each of the 10 experiments, showing the percentage of points in $V$.

   ii. Calculate the fraction of experiments (out of the 10 performed), which resulted in a percentage that was within (plus or minus) 1% of the expected percentage.

   iii. Calculate the sample mean and sample variance of the 10 percentages generated. For the variance, use the Matlab function var(X).

   For your answer to this problem, you do not need to hand in code. Just hand in the bar graph, and the results of the calculations performed above.

   To get you started with Matlab, here is partial code:

```
clear;

%rand() is uniformly distributed random elements
matrix = rand (500 , 10);
```

```
%count the number of entries >=0.7 in each column
vector = sum (matrix >= 0.7) / 500;

%color can be changed using: 'b'(blue) 'r'(red) 'y' (yellow) and so on
bar(vector , 'b');

%set axis
axis([0 11 0 1]);

%set label
xlabel('Column Number');
ylabel('Percentage: %');
legend('Result Bar');
title('Value > = 0.7');

% You need to add code to calculate the quantities specified
% in ii and iii above.
```

(c) Repeat the task in part (b), but this time only generate 100 points in each experiment.

2. In this problem, let $C$ denote the class of concepts, defined on $\mathbb{R}$, represented by an inequality of the form $x \geq \theta$. Again, let $I = [0, 1]$ and let $V$ be the set of points satisfying $x \geq 0.7$.

   In class we showed that a training set size of $N = (1/\epsilon)\ln(2/\delta)$ is sufficient in order to learn target concepts in $C$ in the PAC learning model, provided that given the training set, we choose a consistent hypothesis in $C$. (We actually showed this for concepts of the form $x \leq \theta$, but the same proof applies if we instead consider concepts of the form $x \geq \theta$.)

   Therefore, if we want there to be at least a 97% probability that a consistent concept in $C$ will have at most 1% error, it suffices to have a training set of size $\lceil (1/(.01)) \ln (2/.03) \rceil = 420$. This is an upper bound on the training set size, though, and fewer examples may be enough (especially for particular distributions).

   Suppose the examples in the training set are drawn uniformly at random from $I$, and are labeled according to $V$. Let us see what happens experimentally if we try to learn $V$ by choosing a random sample of 420 points labeled according to $V$, and computing the most-specific concept of the form $x \geq \theta$ that is consistent with the sample (i.e., which has 0 error on the sample).

   Repeat the following experiment 10 times:

   i. Generate 420 random points $x \in \mathbb{R}$ in $I$ as in the previous problem, labeling the points in $V$ as positive, and the points outside $V$ as negative. The resulting set of points $S$ is a sample that is labeled according to $V$.

It should be x >= 0.7
because when it consistent with sample means the sample which contains the data that kind of data satisfy with this hypothesis

ii. Compute hypothesis concept $V_h$, the most-specific concept in $C$ consistent with sample $S$. This hypothesis will be of the form $x \geq \theta_h$, where $\theta_h$ is the smallest positive point in $S$.

iii. Calculate the generalization error (a.k.a. the true error) of $V_h$ with respect to $V$ and the uniform distribution on $I$. This is the probability that a random point drawn uniformly from $I$ will be misclassified by $V_h$, assuming the target is $V$. (Do not estimate the generalization error; you should calculate it exactly.)

iv. Generate another 500 random points in $I$, and calculate the percentage of them that are misclassified by $V_h$. This is an empirical estimate of the generalization error of $V_h$.

Give the results of your 10 experiments in a table with 3 columns, with one row for each experiment. For each experiment, give $V_h$ (by specifying the value of $\theta_h$), the true error you calculated for $V_h$, and the empirical estimate of the generalization error. You do not have to hand in your code.

3. Suppose you go to a medical lab and take a test to determine whether you have a disease that we will call $M$-disease. The lab tells you that your test was positive, but this does not necessarily mean you have $M$-disease. The test returns a correct positive result in 98% of patients with M-disease, and a correct negative result in 97% of the patients without $M$-disease. Only 0.8% of the population has $M$-disease. Use Bayesian reasoning to answer the following questions.

   (a) What was the prior probability that you had $M$-disease, before you took the lab test? 0.8%

   (b) What is the posterior probability that you have $M$-disease, given the positive result of your lab test? 20.85%

   (c) There are two hypotheses in this problem: either you have $M$-disease, or you don't. Which is the MAP hypothesis? Don't have M-disease

   (d) Which of the two hypotheses is the Maximum Likelihood hypothesis?

4. The mean weight of 12-yr-old boys in a certain population was estimated to be normally distributed, with a mean of 116 pounds, and a standard deviation of 42 pounds. The mean weight of 12-yr-old girls in this population was estimated to be normally distributed with a mean of 114 pounds and a standard deviation of 39 pounds. In this population, 40% of the 12-yr-olds are girls, and 60% are boys.

   Given a 12-yr-old child drawn uniformly at random from this population, if the weight of this child is 105 pounds, what is the probability that the child is female? Answer this question using the information about the estimated distributions. Show your work. Use a calculator or program to compute a numerical answer to this problem.