CS 688: Homework 1

Due: Friday Feb 9 by 10:00 PM

Notes:

- Homework is due **by 10:00 PM on the due date.** Remember that you may not use more than 2 late days on any one homework, and you only have a budget of 5 in total.
- Please keep in mind the collaboration policy as specified in the course syllabus. If you discuss questions with others you **must** write their names on your submission, and if you use any outside resources you **must** reference them. **Do not look at each others' writeups, including code.**
- There are 6 problems on 3 pages in this homework.
- Remember that you will be graded on the quality of your report and writing. We may or may not even look at your code. So please write carefully and clearly.
- For your code, please create a single zip file labeled `yourlastname_hw1.zip` and submit it to the "HW1 Code" assignment on Gradescope. You should provide a readme explaining how to run your code and replicate the results in your report.
- All graphs should have clearly labeled axes.

Problems:

1. (25 points) Write a function that flips 1000 different fair coins 10 times each, and reports the proportion of heads for (1) the first coin (ν_1), (2) a randomly chosen coin (ν_r), and (3) the coin that came up heads the least among all of them ($\nu_{\min}$). Let $\mu = 0.5$ be the true proportion of heads in the population. Now, repeat the entire experiment 100,000 times, and plot the histograms of the distributions of $\nu_1, \nu_r, \nu_{\min}$. On another graph, plot estimates for $\Pr(|\nu - \mu| > \epsilon)$ as a function of ϵ, together with the Hoeffding bound. Which coins obey the Hoeffding bound and which do not? Why or why not?

2. (30 points) Consider the following experiment on perceptron learning for random training sets of dimension 10:

   - Generate an 11-dimensional weight vector $\mathbf{w}^*$, where the first dimension is 0 and the other 10 dimensions are sampled independently at random from the uniform $(0, 1)$ distribution (the first dimension will serve as the threshold, and we'll just set it to 0 for convenience).
   - Generate a random training set with 100 examples, where each dimension of each training example is sampled independently at random from the uniform $(-1, 1)$ distribution, and the examples are all classified in accordance with $\mathbf{w}^*$.

- Run the perceptron learning algorithm, starting with the zero weight vector, on the training set you just generated, and keep track of the number of iterations it takes to learn a hypothesis that correctly separates the training data.

Write code to perform the above experiment and then repeat it 1000 times (note that you're generating a new $\mathbf{w}^*$ and a new training set each time). Once you have your code working, plot a histogram of the number of iterations the algorithm takes to learn a linear separator (you should submit this with your writeup). How does the number of iterations compare with the bound on the number of errors we derived in class? Note that this bound will be different for each instantiation of $\mathbf{w}^*$ and the training set, so in order to answer this question, you should analyze the distribution of differences between the bound and the number of iterations. Plot and submit a histogram of the **log** of this difference, and discuss your interpretation of these results.

Can you characterize the situations in which the algorithm takes more iterations to correctly learn a hypothesis that separates the training data? Back up your answer with evidence from your experiments. Hint: You may want to hold $\mathbf{w}^*$ fixed and vary the training set as you try to figure this out.

3. (15 points) In this question, you are asked to prove *Chebyshev's inequality* by proving the following three statements. Each one builds on the prior one, so make sure you are using part (a) to prove part (b) and part (b) to prove part (c).

   a If t is a non-negative random variable, for any $\alpha > 0, \Pr(t \geq \alpha) \leq \mathbb{E}[t]/\alpha$. (Hint: You may want to condition the expectation on the event of interest)

   b If u is any random variable with mean μ and variance σ^2, for any $\alpha > 0$,

   $$\Pr((u-\mu)^2 \geq \alpha) \leq \frac{\sigma^2}{\alpha}$$

   c If $u_1, \ldots, u_n$ are iid random variables, each with mean μ and variance σ^2, and u is the sample mean of $u_1, \ldots, u_n$, then for any $\alpha > 0$,

   $$\Pr((u-\mu)^2 \geq \alpha) \leq \frac{\sigma^2}{n\alpha}$$

   (Hint: You may want to use a standard result about the variance of the sum of several i.i.d. random variables)

4. (10 points) Consider the perceptron in two dimensions, $h(x) = \text{sign}(\mathbf{w} \cdot \mathbf{x})$ where $\mathbf{w} = (w_0, w_1, w_2)^\top$ and $\mathbf{x} = (1, x_1, x_2)^\top$. Here w_0 represents the intercept.

   a Show that the regions on the plane where $h(\mathbf{x}) = +1$ and $h(\mathbf{x}) = -1$ are separated by a line. If we were to instead represent the line by the equation $x_2 = ax_1 + b$, what would the values of a and b be in terms of the elements of $\mathbf{w}$?

   b Draw a picture illustrating the difference between the cases $\mathbf{w} = (1, 2, 3)^\top$ and $\mathbf{w} = (-1, -2, -3)^\top$.

5. (10 points) The log likelihood function of a (univariate) Gaussian distribution is given by $\ln p(x|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln 2\pi$. Derive the estimators $\hat{\mu}$ and $\hat{\sigma^2}$ that maximize this log likelihood function.

6. (10 points) Suppose, as in our usual classification problem, that $y_i \in \{\pm 1\}$, and we are trying to predict the probability that $y = +1$ conditional on $\mathbf{x}$. Show that the maximum likelihood method for any function $h(\mathbf{x})$ with range $[0, 1]$ is equivalent to minimizing

$$\sum_{i=1}^{n} \left(\mathbb{1}[y_i = +1] \ln \frac{1}{h(\mathbf{x}_i)} + \mathbb{1}[y_i = -1] \ln \frac{1}{1 - h(\mathbf{x}_i)} \right)$$

Use this to derive the expression that should be minimized when $h(\mathbf{x}) = \sigma(\mathbf{w} \cdot \mathbf{x})$ where σ is the sigmoid function discussed in class.