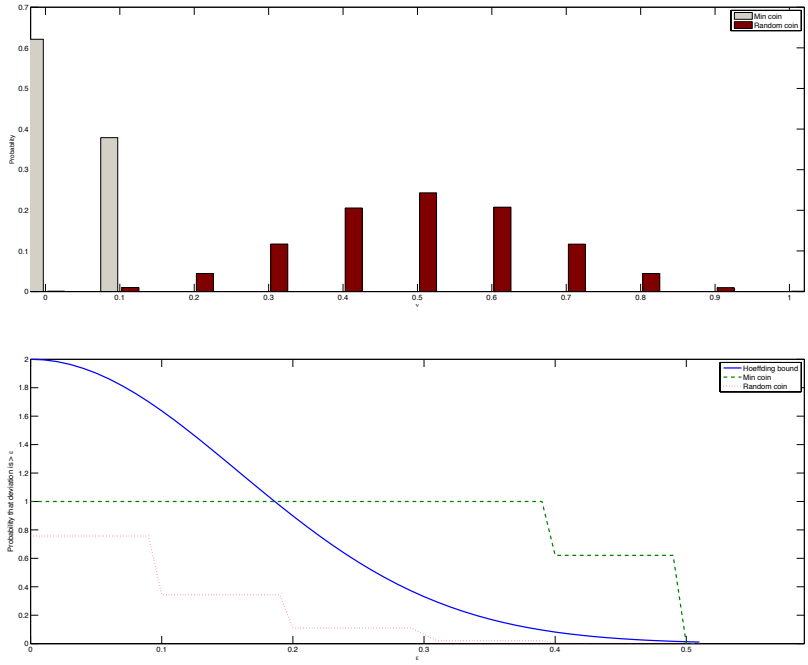
CS 688: Homework 1 Solutions

February 20, 2024

Note: These are not intended to be comprehensive, just to help you see what the answers should be.

1 Your graphs should look something like this (the histograms for the random coin and coin 1 should look almost identical, so we only show one of them here):



The main thing is that coin 1 and the random coin are both the equivalent of hypotheses that are selected before looking at the data, so the Hoeffding bound applies (it's hypothesis verification). On the other hand the data is used to choose the "minimum" coin out of many, so the bound doesn't apply.

2 The graphs should look something like the following. Note: during training, if you train your Perceptron in batches (i.e. train once over the entire dataset) each iteration instead of sampling a random index from the training data, the number of iterations will be two orders of magnitude less (there are 100 data points and you are making 100 "updates" in each "iteration"). Make sure to account for this in the remainder of the analysis.

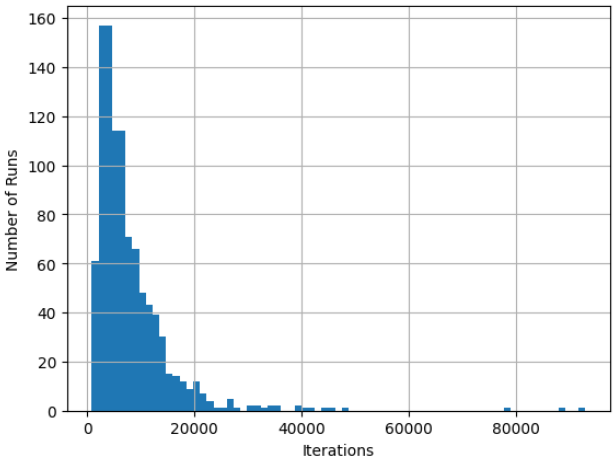


Figure 1: Number of iterations

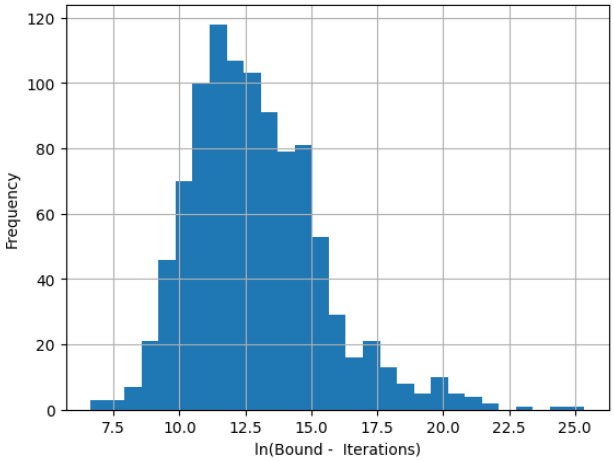
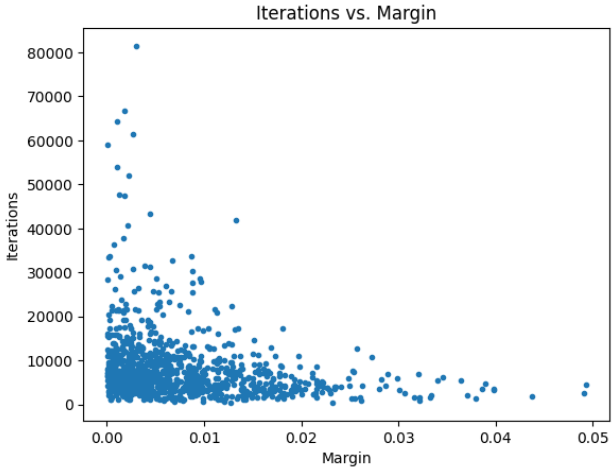


Figure 2: Log of difference between bound and iterations

Situations where the algorithm takes more iterations to learn a hypothesis correctly can be characterized by a small margin in the data. When $\mathbf{w}^*$ is held constant, and the experiment is re-run, you should see something similar to the following figure when you make a scatter-plot of margin vs iterations.



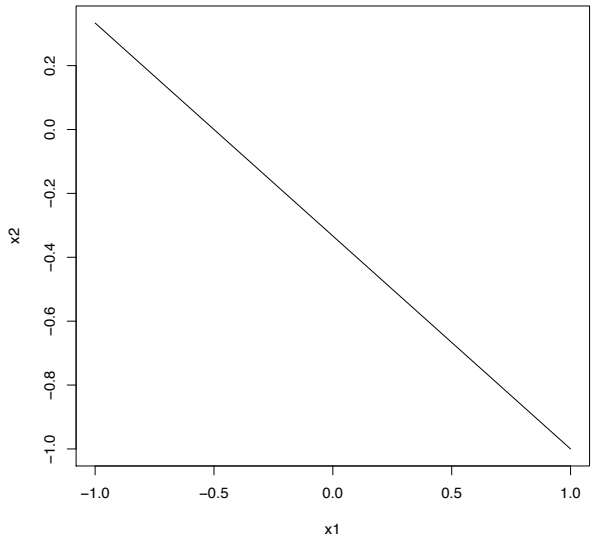
3 For part (a) note that $E(t) = E(t|t < \alpha)\Pr(t < \alpha) + E(t|t \geq \alpha)\Pr(t \geq \alpha) \geq \alpha \Pr(t \geq \alpha)$ (the first term is non-negative and $E(t|t \geq \alpha) \geq \alpha$.

For part (b), let $t = (u - \mu)^2$, which is non-negative. Now, $E(t) = \sigma^2$. Apply part (a) to t:

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

For part (c), since u is the sum of n i.i.d. u_i, each with variance σ^2, the variance of u is σ^2/n (standard result). Now, applying part (b) with the variance σ^2/n gives us the result.

4 For part (a), it is just matter of multiplying **w** and **x** to get a equation of line i.e. $\langle (w_0, w_1, w_2), (1, x_1, x_2) \rangle = w_0 + w_1 x_1 + w_2 x_2$ and since $h(x)$ takes the sign of this equation, this line (rewritten in the form of $y = mx + b$) $x_2 = -\frac{w_1}{w_2} x_1 - \frac{w_0}{w_2}$ divides the plane into two regions making $a = -\frac{w_1}{w_2}$ and $b = -\frac{w_0}{w_2}$. For part (b), the line should look like the following:



For the first case, positive examples will be to the top right of the line, while in the second, they will be to the bottom left.

5 This is just about taking partial derivatives with respect to μ and σ^2 and setting to 0. For $\hat{\mu}$, the partial derivative with respect to μ of the likelihood is $\frac{1}{2\sigma^2} 2 \sum_{i=1}^{n} (x_i - \mu)$ so you will end up with $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$. For the second part, the partial derivative with respect to σ^2 is $\frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 - \frac{n}{2\sigma^2}$ so you will end up with $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$.

6 (This whole thing was basically done in class!)
As in class, maximizing the likelihood is equivalent to minimizing $\sum_{i=1}^{n} \ln \frac{1}{\Pr(y_i | x_i)}$ (safely ignoring the $\frac{1}{n}$ term at the beginning). If $y_i = +1, \Pr(y_i \mid x_i) = h(x_i)$. If $y_i = -1, \Pr(y_i \mid x_i) = 1 - h(x_i)$. Putting these into the above, we see that we need to minimize:

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

For the second part, if $y_i = +1, h(x_i) = \sigma(y_i \mathbf{w}^T \mathbf{x})$. On the other hand, if $y_i = -1, 1 - h(x_i) = 1 - \sigma(\mathbf{w}^T \mathbf{x}) = \sigma(-\mathbf{w}^T \mathbf{x}) = \sigma(y_i \mathbf{w}^T \mathbf{x})$. Therefore, both terms simplify to the same thing, and we end up minimizing the error function $\sum_{i=1}^{n} \ln \frac{1}{\sigma(y_i \mathbf{w}^T \mathbf{x}_i)}$.