Assignment 2

Darwin Ding

September 15, 2016

Exercise 1.8

The probability of picking a marble out of the bag and getting red is 0.9, and we are looking for the probability that we sample 10 random marbles and get 1 or 0 red marbles.

The probability of getting 0 red marbles is $(.1)^{10} = 10^{-10}$, and the probability of getting 1 red marble is $(.1)^9 * .9 * 10$, because we pulled 9 green marbles (with probability .1) and 1 red marble (probability .9), and there were 10 ways to do that.

The probability of doing either of those things is the sum of those two probabilities = $\mathbf{9.1 * 10^{-9}}$

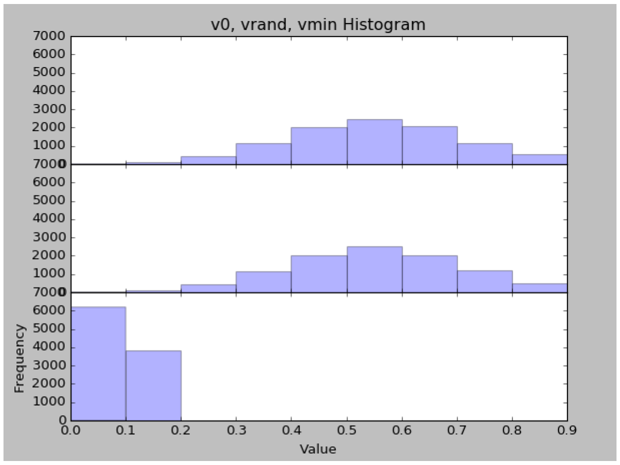
Exercise 1.9

Using Hoeffding's Inequality, $\mu = .9$. Since we're looking for $v \leq .1$, $\epsilon \geq .9 - .1 = .8$. By definition, also, $N = 10$.

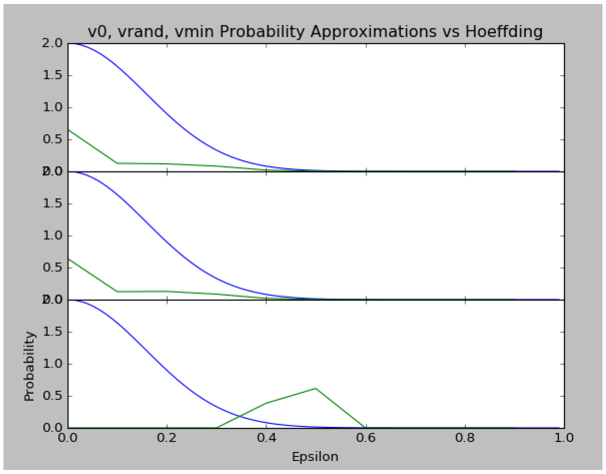
$$
\begin{aligned}
P[|v - \mu| \geq \epsilon] &\leq 2e^{-2\epsilon^2 N} \\
&\leq 2e^{-2*.8^2*10} \\
&\leq \mathbf{5.52 * 10^{-6}}
\end{aligned}
$$

Exercise 1.10

(a) After running the experiment a single time, μ for $v_0 = \mathbf{0.3}$, μ for v_{rand} $= \mathbf{0.5}$ and μ for $v_{min} = \mathbf{0}$

(b) 

The top graph is of v_0, the middle graph is v_{rand}, and the bottom one is v_{min}.

(c) 

Again, the top graph is of v_0, the middle graph is v_{rand}, and the bottom one is v_{min}. The green lines are the calculated estimates, and the blue lines are the plotted Hoeffding bounds.

(d) v_0 and v_{rand} both obey the Hoeffding bound, while v_{min} does not. This makes sense, because v_0 and v_{rand} are essentially randomly selected coins. There's nothing special about the first coin or any random particular coin. The problem arises when you specifically seek out the minimum coin and then try to retroactively perform the Hoeffding bound. Hoeffding's inequality doesn't support this.

(e) If we pretend we don't know the true nature of coin flipping, and that

μ is unknown to us, this problem is very similar to the bins problem. Let's say we are trying to predict μ using our data (the experiments we've performed and their results), and our hypothesis set is all 1000 coins. We then pick out the bins for $[v_0, v_{rand}, v_{min}]$ just to see how well they performed. We have some floating theoretical target function that has $\mu = 0.5$, red/green marbles represent heads/tails, and we're trying to find the true μ.

If we fix our hypothesis set and then perform the experiment, we find that the Hoeffding bounds work simply because they work for $[v_0, v_{rand}]$. We cannot, however, claim that the bounds don't work simply because they retroactively don't work for v_{min}.

Exercise 1.11

(a) No, not necessarily. If $p = 0.5$, then even if E_{in} is approximately equal to E_{out}, $E_{in} = 0.5$ which is just as good as random.

(b) Yes. Just because the smart algorithm has the smallest in-sample error (even if it's 0) doesn't mean there's a chance that the out-of-sample data is completely different from the in-sample. In this case, all of the points out of D might be -1, and the crazy learning algorithm will have worked better.

(c) If $p = 0.9$, S will pick the "correct" hypothesis unless $E_{in}(h_1) > 0.5$. The probability of that is the probability that $\epsilon > .4$ ($E_{out}(h_1) = .1$). M = 2, so this probability is bounded by $2 * 2 * e^{-2\epsilon^2 N} = .0013$. Therefore the probability that S correctly hypothesizes from the set is at least 1 - .0013 = **0.99866**

(d) For any p, $E_{in}(h_1) = 1 - p$. $E_{out}(h_1)$ needs to be on the opposite 'side' of E_{in} with probability greater than 0.5 for the C algorithm to work. In other words, $|E_{out} - E_{in}| > \epsilon$, and $\epsilon = |p - 0.5|$.

Solving for $2 * 2 * e^{-2(p-0.5)^2 * 25} > 0.5$ gets us $\mathbf{0.296 < p < 0.704}$

Exercise 1.12

The best you can provide is **(c)**. Depending on the problem, you may need to have a very large hypothesis set, which would decrease your out-of-bound prediction accuracy. In that scenario, you may have to announce that you have failed. However, if you find that E_{out} is not close to 0.5, then you can announce that you have found a function that approximates well out-of-sample.

Problem 1.3

(a) For separable data, the optimal set of weights will split the data such that for every point, ${w^*}^T x_n$ has the same sign as y_n. If that is not true, then either the data is not separable or this is not the optimal set of weights. Since these all have the same sign, $\rho = {w^*}^T x_n * y_n$ will always be positive for all n.

(b) According to the update rule, $w(t+1) = w(t) + y(t)x(t)$. Let's assume that we have some misclassified x(m) point with a corresponding y(m).

$$\begin{aligned} w^T(t)w^* &= {w^*}^T(w(t-1) + x(m)y(m)) \\ &= {w^*}^T w(t-1) + {w^*}^T x(m)y(m) \end{aligned}$$

${w^*}^T x(m)y(m)$ is positive for all n, so ${w^*}^T x(m)y(m) \geq \rho$, since ρ is the smallest in magnitude of ${w^*}^T x(n)y(n)$ for all n.

Therefore, $w^T(t)w^* = {w^*}^T w(t-1) + {w^*}^T x(m)y(m) \geq {w^*}^T w(t-1) + \rho$.

Base Case
At t = 0, $w^T(t)$ is a vector of all 0s, and so both sides of the inequality are equal at 0.

Inductive Case
At t = n, assume $w^T(n)w^* \geq t\rho$. At t = n+1, we've established in the previous portion that $w^T(n+1)w^* \geq {w^*}^T w(t) + \rho \geq t\rho + \rho$.

Therefore, $w^T(n+1)w^* \geq (t+1)\rho$ and the conclusion works.

(c)

$$||w(t)^2|| = ||(w(t-1) + x(t-1)y(t-1))^2||$$
$$= ||w(t-1)^2 + 2w^T(t-1)x(t-1)y(t-1) + x^2(t-1)y^2(t-1)||$$
$$= ||w(t-1)^2 + 2w^T(t-1)x(t-1)y(t-1) + x^2(t-1)||$$

By definition, $||a|| + ||b|| \geq ||a+b||$. We know that $w^T(t-1)x(t-1)y(t-1) < 0$ simply because x(t-1) was selected because the PLA misclassified it at time t-1. Therefore:

$$||w(t-1)^2 + 2w^T(t-1)x(t-1)y(t-1) + x^2(t-1)||$$
$$< ||w(t-1)^2 + x^2(t-1)||$$
$$\leq ||w(t-1)||^2 + ||x(t-1)||^2$$

(d) **Base Case**
At t = 0, $||w(t)||^2 = 0$, so both sides of the inequality are 0 again.

Inductive Case
At t = n, assume $||w(t)||^2 \leq tR^2$. At t = n+1, we've established in the previous portion that $||w(t+1)||^2 \leq ||w(t)||^2 + ||x(t)||^2$. We also know that by definition, $x(t) \leq R$ (meaning $||x(t)||^2 \leq R^2$)

Therefore, $||w(t+1)||^2 \leq ||w(t)||^2 + ||x(t)||^2 \leq tR^2 + R^2 = (t+1)R^2$ and the conclusion works.

(e) We know from (b), $w^T(t)w^* \geq t\rho$, and we know from (d), $||w(t)||^2 \leq tR^2$, thus $||w(t)|| \leq \sqrt{t}R$. This can be further transformed as follows:

$$||w(t)|| \leq \sqrt{t}R$$
$$1/||w(t)|| \geq 1/(\sqrt{t}R)$$

Multiplying these two inequalities together, we find that $w^T(t)w^*/||w(t)|| \geq t\rho/(\sqrt{t}R) = \sqrt{t}\rho/R$. Squaring both sides, we get:

$$w^T(t)w^*/||w(t)|| \geq \sqrt{t}\rho/R$$
$$(w^T(t) * w^*)^2/||w(t)||^2 \geq t\rho^2/R^2$$
$$\frac{R^2(w^T(t) * w^*)^2}{\rho^2||w(t)||^2} \geq t$$

One key thing to note is that due to the definition of the dot product, where $A \cdot B = ||A||||B||cos(\theta)$, $A \cdot B = A^T B \leq ||A||||B||$ since $-1 \leq cos(\theta) \leq 1$. Therefore, we can further transform our equation as follows:

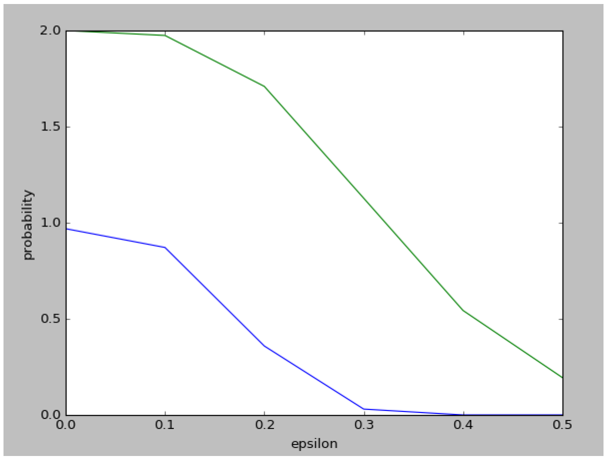
$$\frac{R^2(||w(t)|| * ||w^*||)^2}{\rho^2||w(t)||^2} \geq \frac{R^2(w^T(t) * w^*)^2}{\rho^2||w(t)||^2} \geq t$$
$$\frac{R^2||w^*||^2}{\rho^2} \geq t$$

Problem 1.7

(a) For 1 coin, the probability of getting no heads is equal to $\binom{10}{0}(1 - .05)^{10} = \mathbf{0.598}$

For 1000 coins, we can use the previous result. P[at least 1 coin gets no heads] = 1 - P[all coins get at least 1 head] $= 1 - (1 - .598)^{1000}$ which is approximately **1**

We can continue this formula for 1,000,000, which will pretty much give us the same value: approximately **1**

(b) 

The green line represents the Hoeffding bounds (calculated using P[coin 1 exceeds ϵ] + P[coin 2 exceeds ϵ] - P[coin 1 ...]P[coin 2 ...]), and the blue line represents the raw calculation from iterating and using the binomial expansion for each case.