Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Solution assignment 1
Due: February 22, 2010

A. PAC Learning

- Let $X = \mathbb{R}^2$ with orthonormal basis (e_1, e_2) and consider the set of concepts defined by the area inside a right triangle ABC with two sides parallel to the axes, with $\overrightarrow{AB}/AB = e_1$ and $\overrightarrow{AC}/AC = e_2$, and $AB/AC = \alpha$ for some positive real $\alpha \in \mathbb{R}_+$. Show, using similar methods to those used in the lecture slides for the axis-aligned rectangles, that this class can be (ϵ, δ)-PAC-learned from training data of size $m \geq (3/\epsilon)\log(3/\delta)$.

  As in the case of axis-aligned rectangles, consider three regions r_1, r_2, r_3, along the sides of the target concept as indicated in Figure 1. Note that the triangle formed by the points $A", B", C"$ is similar to ABC (same angles) since $A"B"$ must be parallel to AB, and similarly for the other sides.

  Assume that $\Pr[ABC] > \epsilon$, otherwise the statement would be trivial. Consider a triangle $A'B'C'$ similar to ABC and consistent with the training sample and such that it meets all three regions r_1, r_2, r_3.

  Since it meets r_1, the line $A'B'$ must be below $A"B"$. Since it meets r_2 and r_3, A' must be in r_2 and B' in r_3 (see Figure 1). Now, since the angle $\widehat{A'B'C'}$ is equal to $\widehat{A"B"C"}$, C' must be necessarily above $C"$. This implies that triangle $A'B'C'$ contains $A"B"C"$ and thus $error(A'B'C') \leq \epsilon$.

$$error(A'B'C') > \epsilon \implies \exists i \in \{1,2,3\} \colon A'B'C' \cap r_i = \emptyset.$$

  Thus, by the union bound,

$$\Pr[error(A'B'C') > \epsilon] \leq \sum_{i=1}^{3} \Pr[A'B'C' \cap r_i = \emptyset] \leq 3(1-\epsilon/3)^m \leq 3e^{-3m\epsilon}.$$

  Setting δ to match the right-hand side gives the sample complexity $m \geq \frac{3}{\epsilon}\log\frac{3}{\delta}$.

B. Biased coins [based on theorem of (Anthony and Bartlett, 1999)]

- Professor Moent has two coins in his pocket, coin x_A and coin x_B, both slightly biased as follows:

$$\Pr[x_A = 0] = 1/2 - \epsilon/2 \quad \Pr[x_B = 0] = 1/2 + \epsilon/2,$$

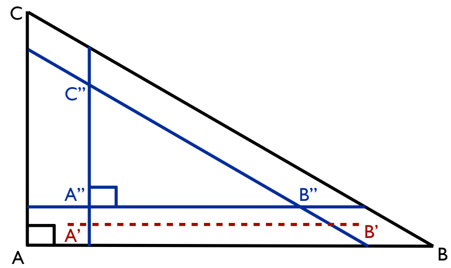

Figure 1: Rectangle triangles.

where $0 < \epsilon < 1$ is a small positive number and where 0 denotes head and 1 tail. He likes to play the following game with his students. He randomly picks a coin $x \in \{x_A, x_B\}$ from his pocket (uniformly), tosses it m times, then reveals the sequence of 0s and 1s he obtained and asks which coin was tossed.

The goal of this exercise is to determine how large m needs to be for a student's coin prediction error to be at most $\delta > 0$.

1. Let S be a sample of size m. Professor Moent's best student, Oskar, plays according to the decision rule $f_o \colon \{0,1\}^m \to \{x_A, x_B\}$ defined by $f_o(S) = x_A$ iff $N(S) < m/2$, where $N(S)$ is the number of 0's in sample S.

   Suppose m is even, then show that

$$error(f_o) \geq \frac{1}{2} \Pr\left[N(S) \geq \frac{m}{2} \Big| x = x_A\right]. \tag{1}$$

By definition of the error of Oskar's prediction rule,

$$\begin{aligned}
error(f_o) &= \Pr[f_o(S) \neq x] \\
&= \Pr[f_o(S) = x_A \wedge x = x_B] + \Pr[f_o(S) = x_B \wedge x = x_A] \\
&= \Pr\left[N(S) < \frac{m}{2} \Big| x = x_B\right] \Pr[x = x_B] + \\
&\quad \Pr\left[N(S) \geq \frac{m}{2} \Big| x = x_A\right] \Pr[x = x_A] \\
&= \frac{1}{2} \Pr[N(S) < \frac{m}{2} \Big| x = x_B] + \frac{1}{2} \Pr\left[N(S) \geq \frac{m}{2} \Big| x = x_A\right] \\
&\geq \frac{1}{2} \Pr\left[N(S) \geq \frac{m}{2} \Big| x = x_A\right].
\end{aligned}$$

2. Assuming m even, use the inequalities given in the appendix to show that

$$error(f_o) > \frac{1}{4}\Big[1 - \Big[1 - e^{-\frac{m\epsilon^2}{1-\epsilon^2}}\Big]^{\frac{1}{2}}\Big]. \tag{2}$$

Note that $\Pr[N(S) \geq \frac{m}{2} | x = x_A] = \Pr[B(m,p) \geq k]$, *with* $p = 1/2 - \epsilon/2$, $k = \frac{m}{2}$, *and* $mp \leq k \leq m(1-p)$. *Thus, by the binomial inequality of the Appendix,*

$$error(f_o) \geq \frac{1}{2}\Pr\left[N \geq \frac{m\epsilon/2}{\sqrt{1/4(1-\epsilon^2)m}}\right] = \frac{1}{2}\Pr\left[N \geq \frac{\sqrt{m}\epsilon}{\sqrt{1-\epsilon^2}}\right].$$

Using the second inequality of the Appendix, we now obtain

$$error(f_o) \geq \frac{1}{4}\Big(1 - \sqrt{1 - e^{-u^2}}\Big),$$

with $u = \frac{\sqrt{m}\epsilon}{\sqrt{1-\epsilon^2}}$, *which coincides with (2).*

3. Argue that if m is odd, the probability can be lower bounded by the one for $m+1$ and conclude that for both odd and even m,

$$error(f_o) > \frac{1}{4}\Big[1 - \Big[1 - e^{-\frac{2\lceil m/2\rceil\epsilon^2}{1-\epsilon^2}}\Big]^{\frac{1}{2}}\Big]. \tag{3}$$

If m *is odd, since* $\Pr\left[N(S) \geq \frac{m}{2}\Big| x = x_A\right] \geq \Pr\left[N(S) \geq \frac{m+1}{2}\Big| x = x_A\right]$, *we can use the lower bound*

$$error(f_o) \geq \frac{1}{2}\Pr\left[N(S) \geq \frac{m+1}{2}\Big| x = x_A\right].$$

Thus, in both cases we can use the lower bound expression with $\lceil m/2\rceil$ *instead of* $m/2$.

4. Using this bound, how large must m be if Oskar's error is at most δ, where $0 < \delta < 1/4$. What is the asymptotic behavior of this lower bound as a function of ϵ?

If $error(f_o)$ *is at most* δ, *then* $\frac{1}{4}\Big[1 - \Big[1 - e^{-\frac{2\lceil m/2\rceil\epsilon^2}{1-\epsilon^2}}\Big]^{\frac{1}{2}}\Big] < \delta$, *which gives*

$$e^{-\frac{2\lceil m/2\rceil\epsilon^2}{1-\epsilon^2}} < 1 - (1-4\delta)^2 = 4\delta(2-4\delta) = 8\delta(1-2\delta),$$

and

$$m > 2\left\lceil \frac{1-\epsilon^2}{2\epsilon^2}\log\frac{1}{8\delta(1-2\delta)}\right\rceil.$$

The lower bound varies as $\frac{1}{\epsilon^2}$.

5. Show that no decision rule $f\colon \{0,1\}^m \to \{x_a, x_B\}$ can do better than Oskar's rule f_o. Conclude that the lower bound of the previous question applies to all rules.

   Let f be an arbitrary rule and denote by F_A the set of samples for which $f(S) = x_A$ and by F_B the complement. Then, by definition of the error,

$$\begin{aligned}
error(f) &= \sum_{S \in F_A} \Pr[S \wedge x_B] + \sum_{S \in F_B} \Pr[S \wedge x_A] \\
&= \frac{1}{2} \sum_{S \in F_A} \Pr[S|x_B] + \frac{1}{2} \sum_{S \in F_B} \Pr[S|x_A] \\
&= \frac{1}{2} \sum_{\substack{S \in F_A \\ N(S) < m/2}} \Pr[S|x_B] + \frac{1}{2} \sum_{\substack{S \in F_A \\ N(S) \geq m/2}} \Pr[S|x_B] + \\
&\quad \frac{1}{2} \sum_{\substack{S \in F_B \\ N(S) < m/2}} \Pr[S|x_A] + \frac{1}{2} \sum_{\substack{S \in F_B \\ N(S) \geq m/2}} \Pr[S|x_A].
\end{aligned}$$

   Now, if $N(S) \geq m/2$, clearly $\Pr[S|x_B] \geq \Pr[S|x_A]$. Similarly, if $N(S) < m/2$, clearly $\Pr[S|x_A] \geq \Pr[S|x_B]$. In view of these inequalities, $error(f)$ can be lower bounded as follows

$$\begin{aligned}
error(f) &\geq \frac{1}{2} \sum_{\substack{S \in F_A \\ N(S) < m/2}} \Pr[S|x_B] + \frac{1}{2} \sum_{\substack{S \in F_A \\ N(S) \geq m/2}} \Pr[S|x_A] + \\
&\quad \frac{1}{2} \sum_{\substack{S \in F_B \\ N(S) < m/2}} \Pr[S|x_B] + \frac{1}{2} \sum_{\substack{S \in F_B \\ N(S) \geq m/2}} \Pr[S|x_A] \\
&= \frac{1}{2} \sum_{S\colon N(S) < m/2} \Pr[S|x_B] + \frac{1}{2} \sum_{S\colon N(S) \geq m/2} \Pr[S|x_A] \\
&= error(f_o).
\end{aligned}$$

   Oskar's rule is known as the maximum likelihood *solution.*

Appendix

0.1 Binomial inequality

Let B be a binomial (m, p) random variable with with $p \leq 1/2$. Then, for $mp \leq k \leq m(1-p)$,

$$\Pr[B \geq k] \geq \Pr\left[N \geq \frac{k - mp}{\sqrt{mp(1-p)}}\right], \tag{4}$$

where N is in standard normal form.

0.2 Tail bound

If N is a random variable following the standard normal distribution, then for $u > 0$,

$$\Pr[N \geq u] \geq \frac{1}{2}\Big(1 - \sqrt{1 - e^{-u^2}}\Big). \tag{5}$$