Mehryar Mohri
Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 1 - solution
Credit: Ashish Rastogi, Afshin Rostamizadeh
Ameet Talwalkar, and Eugene Weinstein.

1. **Probability Review** [30 points]:

   (a) [10 points] Imagine you are given one fair die, and you need to decide which task is harder: (i) guessing the value of one die toss or (ii) tossing the die twice and getting the same value twice. Given that the die is fair (every side has weight 1/6), does event (i) have a greater chance of success or event (ii), or do they have the same probability of success? Make sure to give justification.

   Solution: Successfully guessing the outcome of either event is equally likely. Clearly, guessing the value of a single fair toss is 1/6. To see that the probability of rolling the same value twice, let X_1 denote the outcome of the first toss and X_2 denote the value of the second toss. Then, we are interested in the value of $\Pr(X_1 = X_2) = \sum_{i=1}^{6} \Pr(X_1 = i \wedge X_2 = i)$. Notice each toss is independent and identical, so we can write

$$\begin{aligned}\sum_{i=1}^{6} \Pr(X_1 = i \wedge X_2 = i) &= \sum_{i=1}^{6} \Pr(X_1 = i) \Pr(X_2 = i) \\ &= \sum_{i=1}^{6} 1/6 \cdot 1/6 \\ &= 1/6\end{aligned}$$

   (b) [5 points] We will now generalize this result to n-sided dice with any (possibly non-uniform) distribution. First prove the following useful fact, for any $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that $\sum_i \alpha_i = 1$, the following holds,

$$0 \leq \sum_{i=1}^{n} (\alpha_i - 1/n)^2 = \sum_{i=1}^{n} \alpha_i^2 - 1/n$$

Solution: The inequality is true, because a sum of squares is always positive. To show the equality, we simply expand the terms and use the fact $\sum_i \alpha_i = 1$.

$$\begin{aligned}\sum_{i=1}^{n}(\alpha_i - 1/n)^2 &= \sum_{i=1}^{n}(\alpha_i^2 - \frac{2}{n}\alpha_i + 1/n^2) \\ &= \sum_{i=1}^{n}\alpha_i^2 - \frac{2}{n}\sum_{i=1}^{n}\alpha_i + \sum_{i=1}^{n}1/n^2 \\ &= \sum_{i=1}^{n}\alpha_i^2 - 2/n + 1/n \\ &= \sum_{i=1}^{n}\alpha_i^2 - 1/n\end{aligned}$$

(c) [15 points] Let X_1 be the value of the first toss, and X_2 be the value of the second toss. Show that $\Pr(X_1 = X_2) \geq 1/n$ (hint: use part b). For what distribution is the inequality tight?

Solution: Generalizing part (a) in a straight-forward manner, we get

$$\begin{aligned}\Pr(X_1 = X_2) &= \sum_{i=1}^{n}\Pr(X_1 = i \wedge X_2 = i) \\ &= \sum_{i=1}^{n}\Pr(X_1 = i)\Pr(X_2 = i) \quad \text{(independent)} \\ &= \sum_{i=1}^{n}\Pr(X_1 = i)^2 \quad \text{(and identical)}\end{aligned}$$

Notice that, $\sum_{i=1}^{n}\Pr(X_1 = i) = 1$ by definition, so we can think of $\Pr(X_1 = i) = \alpha_i$. From part (b), we know that $\sum_{i=1}^{n}\alpha_i^2 \geq 1/n$. As seen in part (a), the uniform distribution ($\alpha_i = 1/n$) achieves equality.

2. **Concentration Bounds** [30 points]:

(a) [10 points] Given a sample of m bounded points $X = (x_1, x_2, \ldots, x_m)$, $\forall i, |x_i| \leq M$, define the function

$$f(X) = \frac{1}{m}\sum_i x_i.$$

Can you give a bound on the probability $\Pr[|f(X) - \mathrm{E}[f(X)] \geq \epsilon]$?

Solution: This is simply a straight-forward application of Hoeffding's inequality. We can think of our function f as the sum of random variables x_i/m, and we know the value of each variable is bounded by $2M/m$. Hoeffding's inequality give the following bound,

$$\begin{aligned}
\Pr[|f(X) - \mathrm{E}[[]f(X)]| > \epsilon] &\leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m (2M/m)^2}\right) \\
&= 2\exp\left(\frac{-2\epsilon^2 m}{4M^2}\right)
\end{aligned}$$

(b) [10 points] Let X and X' be two sets of size m that differ in exactly one point. That is, $|X \cap X'| = m - 1$. We say a function h is *stable* if for all such X, X', $|h(X) - h(X')| \leq g(m)$ for some decreasing function g. How quickly does g need to decrease as a function of m in order for McDiarmid's inequality to provide a bound on the event $\Pr[|h(X) - \mathrm{E}[h(X)]| \geq \epsilon]$ that converges to zero as $m \to \infty$?

Solution: In order for McDiarmid's inequality to converge, we need $g(m) \in o(1/\sqrt{m})$. In the case $g(m) = 1/m^{1/2+\delta}$, we can apply McDiarmid's inequality with each $c_i = 1/m^{1/2+\delta}$,

$$\begin{aligned}
\Pr[|h(X) - E_X[h(X)]| \geq \epsilon] &\leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^m (m^{-1/2} m^{-\delta})^2}\right) \\
&= 2\exp\left(-2\epsilon^2 m^{2\delta}\right)
\end{aligned}$$

Clearly we need $\delta > 0$ in order for the bound to converge to zero as m tends to infinity.

(c) [10 points] Is the function f from part (a) stable (still assuming the bound $|x_i| \leq M, \forall i$)? Will McDiarmid's inequality provide provide a convergent bound? If so give the bound. Now define the function $f'(X) = \max(X)$, is f' stable? Can you give a bound with McDiarmid's inequality?

Solution: The function from part (a) is stable, with $g(m) = 2M/m$ (changing a single bounded point, will change the average by at most $2M/m$). Indeed, from part (b), we can reason that McDiarmid's inequality, with $c_i = 2M/m$ will give a convergent bound. In fact, in this case, the bound is exactly the same as the one given by Hoeffding's inequality. Here we see that McDiarmid's inequality generalized Hoeffding's.

The function f' is not stable, the only bound that we can give is $g(m) \leq M$ (without assuming anything else about the distribution). We cannot get a useful bound with McDiarmid's inequality.

3. **PAC Learning** [40 points + 20 points]: Here we will consider an alternative PAC learning scenario, called the two-oracle model. Imagine you are given the ability to explicitly ask for a positive or negative sample, which are drawn from different distributions D_+ and D_- respectively. A concept is efficiently PAC-learnable if there exists an algorithm L that can generate a hypothesis h, such that $\Pr_{x\sim D_+}[h(x) = 0] \leq \epsilon$ and $\Pr_{x\sim D_-}[h(x) = 1] \leq \epsilon$ with confidence $(1-\delta)$, after sampling $m = \text{poly}(1/\epsilon, 1/\delta)$ points.

   (a) [40 points] Show that if a problem is efficiently PAC-learnable in the classic sense, it is also always efficiently PAC-learnable in the two-oracle model.

   Solution: Let c be the true concept, then notice that

$$\begin{aligned}
\text{error}(h) &= \Pr_x[h(x) \neq c(x)] \\
&= \Pr_x[h(x) = 1 \wedge c(x) = 0] + \Pr_x[h(x) = 0 \wedge c(x) = 1] \\
&= \Pr_x[h(x) = 1 | c(x) = 0] \Pr_x[c(x) = 0] + \\
&\quad \Pr_x[h(x) = 0 | c(x) = 1] \Pr_x[c(x) = 1].
\end{aligned}$$

   Let $0 < \epsilon < 1/2$, then by assumption we know there exists an algorithm L that will efficiently produce a hypothesis h, such that $\text{error}(h) \leq \epsilon/2$ with confidence $1 - \delta$. From the above series of equalities this implies,

$$\Pr_x[h(x) = 1 | c(x) = 0] \Pr_x[c(x) = 0] + \Pr_x[h(x) = 0 | c(x) = 1] \Pr_x[c(x) = 1] \leq \epsilon/2. \quad (1)$$

Thus, in the two-oracle model, we can simulate the classic scenario by simply treating points from either distribution D_+ or D_- as coming from the same underlying distribution. If we draw points uniformly from the negative and positive oracle we will have $\Pr_x[c(x) = 0] = \Pr_x[c(x) = 1] = 1/2$, which would imply $\Pr_x[h(x) = 1|c(x) = 0] = \Pr_{x \sim D_-}[h(x) = 1] \leq \epsilon$, and similarly $\Pr_x[h(x) = 0|c(x) = 1] = \Pr_{x \sim D_+}[h(x) = 0] \leq \epsilon$.

(b) [20 points] (Bonus) Show that the reverse direction is also true.