Homework 2

Due: March 2, 2018 at 7:00 PM

Written Questions

Problem 1

(5 points)

Let $x_1, x_2, ..., x_m$ be an i.i.d. (independent and identically distributed) sample drawn from distribution $B(p)$ where $B(p)$ denotes a Bernoulli distribution. Specifically,

$$\mathbb{P}(B = 1) = p, \quad \mathbb{P}(B = 0) = 1 - p.$$

Suppose p is an unknown parameter and we estimate it via:

$$\widehat{p} = \frac{1}{m} \sum_{i=1}^{m} (x_i).$$

Show that $\widehat{p}$ is an unbiased estimator for p. Recall that an estimator is unbiased if its expected value over all possible samples is equal to the parameter it is estimating. *Note:* In lecture, we showed this estimator is unbiased for a sample size of 3. For this question, we are asking you to generalize the argument to a sample size of m.

Solution: We want to show that $\mathbb{E}[\hat{p}] = p$ for an arbitrary m.

We know that $\mathbb{E}[x_i] = 1 \cdot p + 0 \cdot (1 - p) = p$ for any sample x_i, so using linearity of expectation:

$$\begin{aligned} \mathbb{E}[\hat{p}] &= \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} x_i\right] \\ &= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[x_i] \\ &= \frac{1}{m} \sum_{i=1}^{m} p \\ &= p \end{aligned}$$

Thus $\mathbb{E}[\hat{p}] = p$, so our estimator is unbiased.

One could also use the binomial theorem, though it is more complicated:

$$\begin{aligned}
\mathbb{E}[\hat{p}] &= \mathbb{E}\left[\frac{\text{\# of 1's}}{\text{\# of samples}}\right] \\
&= \sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} \frac{k}{m} \\
&= \sum_{k=1}^{m} \binom{m}{k} p^k (1-p)^{m-k} \frac{k}{m} \qquad \text{(since we get 0 when k=0)} \\
&= \sum_{k=1}^{m} \binom{m-1}{k-1} p^k (1-p)^{m-k} \\
&= p \cdot \sum_{k=1}^{m} \binom{m-1}{k-1} p^{k-1} (1-p)^{(m-1)-(k-1)} \\
&= p \cdot \sum_{k=0}^{m-1} \binom{m-1}{k} p^{k} (1-p)^{(m-1)-k} \\
&= p \cdot (p + (1-p))^{m-1} \qquad \text{(using the binomial theorem)} \\
&= p
\end{aligned}$$

Problem 2: Agnostic PAC Learning

(25 points)

Previously, we looked at PAC learning under the assumption that the true hypothesis was a function within our set of hypotheses H—the *realizable case*. However, this assumption does not always hold. In some cases, the true hypothesis is a function $f \notin H$—the *unrealizable case*. Learning in the unrealizable case is also called **agnostic learning**. We will examine how PAC learning differs in this scenario.

Hoeffding's inequality can give us a bound on sample size for the unrealizable case in which we have a finite set of hypotheses H. The true function for labeling data is f. Suppose we are labelling data x generated from distribution D. Define the *expected error* of a hypothesis, $\mathrm{err}_D(h)$, as the expected proportion of data incorrectly labelled by the hypothesis h, written as:

$$\mathrm{err}_D(h) = \mathbb{E}_{x \sim D}[f(x) \neq h(x)].$$

We have a sample S. Define the *sampling error* of a hypothesis, $\mathrm{err}_S(h)$, as the proportion of data from the sample S incorrectly labelled by hypothesis h. It may be written as:

$$\mathrm{err}_S(h) = \frac{1}{|S|} \sum_{(x,y) \in S} [y \neq h(x)].$$

Define $\mathrm{ERM}(H) = \mathrm{argmin}_{h \in H}\, \mathrm{err}_S(h)$ and $\mathrm{RM}(H) = \mathrm{argmin}_{h \in H}\, \mathrm{err}_D(h)$ as the empirical risk minimizing hypothesis and the risk minimizing hypothesis, respectively. In the PAC setting, we content ourselves with an algorithm that, with probability at least $1 - \delta$, returns a hypothesis $\widehat{h}$ whose error over the distribution D is within ϵ of that of the risk minimizing solution, or:

$$|\mathrm{err}_D(\widehat{h}) - \mathrm{err}_D(\mathrm{RM}(H))| \leq \epsilon.$$

Note: We are considering PAC learning on a binary dataset.

a. First, consider the realizable case, $f \in H$. What is the PAC learning bound for $\widehat{h} = \mathrm{ERM}(H)$? Express the sample size bound in terms of ϵ, δ, and $|H|$.

b. Suppose we're stuck with agnostic learning such that $f \notin H$. Our best hypothesis is $\widehat{h} = \text{ERM}(H)$. Use Hoeffding's inequality to show that, given ϵ_1 and δ_1, there's an m such that sampling $S = \{x_1, ..., x_m\} \sim D$ gives us $|\text{err}_S(h) - \text{err}_D(h)| \leq \epsilon_1$, with probability at least $1 - \delta_1$, for some h.

c. What happens if we use a sample of size m to evaluate *all* the hypotheses in H? In particular, we want it to be simultaneously true that, for all $h \in H$, $|\text{err}_S(h) - \text{err}_D(h)| \leq \epsilon_1$ with probability $1 - \delta$. Write an upper bound for the true failure probability in terms of δ_1 using the union bound. Then, express δ in terms of δ_1 so that with probability $1 - \delta$ it is true that $|\text{err}_S(h) - \text{err}_D(h)| \leq \epsilon_1$ for all $h \in H$.

d. Now we know that our error estimates for all $h \in H$ are within ϵ_1 of their true errors (with high probability). If we pick $\widehat{h} = \text{ERM}(H)$, how far might $\widehat{h}$ be from $\text{RM}(H)$? Call that ϵ and write ϵ in terms of ϵ_1. That is, find ϵ in terms of ϵ_1 such that $|\text{err}_D(\widehat{h}) - \text{err}_D(\text{RM}(H))| \leq \epsilon$.

e. Putting it all together, define m in terms of ϵ and δ so that, with probability at least $1 - \delta$ (using the previously computed bound), $|\text{err}_D(\text{ERM}(H)) - \text{err}_D(\text{RM}(H))| \leq \epsilon$.

Solution:

1. Using the theorem given in class, this is just want our sample size m to satisfy $m \geq \dfrac{\log\left(\frac{|H|}{\delta}\right)}{\epsilon}$

2. Hoeffding says that

$$\mathbb{P}(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i - \mu\right| > \epsilon) \leq 2\exp\left(-2m\epsilon^2/(a-b)^2\right)$$

where a and b are bounds on the range of values our samples can take. For our 0-1 loss function, these values are 1 and 0 respectively. Substituting in $\text{err}_S(h)$, $\text{err}_D(h)$, ϵ_1, and δ_1, and using algebra, we get

$$\begin{aligned}
\delta_1 &\geq 2\exp\left(-2m\epsilon_1^2\right) \\
\implies \log(\frac{\delta_1}{2}) &\geq -2m\epsilon_1^2 \\
\implies \frac{-\log(\frac{2}{\delta_1})}{2\epsilon_1^2} &\geq -m \\
\implies \frac{\log(\frac{2}{\delta_1})}{2\epsilon_1^2} &\leq m
\end{aligned}$$

3. Consider how $P(A \cup B) = P(A) + P(B) - P(A \cap B) \leq P(A) + P(B)$. Using the last expression, if the probability that $\text{err}_S(h)$ differs significantly from $\text{err}_D(h)$ is δ_1 for all h, then setting $\delta \geq |H|\delta_1$ is sufficient to guarantee that the probability that all are correct is at least $1 - \delta$.

4.

$$\begin{aligned}
|\text{err}_D(\hat{h}) - \text{err}_D(RM(H))| &= \text{err}_D(\hat{h}) - \text{err}_D(RM(H)) \qquad \text{(because RM(H) is the minimizer of this quantity)} \\
&\leq \text{err}_S(\hat{h}) + \epsilon_1 - \text{err}_D(RM(H)) \qquad \text{(with high probability)} \\
&\leq \text{err}_S(\hat{h}) + \epsilon_1 - \text{err}_S(RM(H)) + \epsilon_1 \qquad \text{(again applying bound from previous part)} \\
&\leq \text{err}_S(\hat{h}) + \epsilon_1 - \text{err}_S(\hat{h}) + \epsilon_1 \qquad \text{(because } \hat{h} \text{ is minimizer of this err}_S(\cdot)\text{)} \\
&\leq 2\epsilon_1
\end{aligned}$$

5. Plugging into our answer from part 2:

$$m \geq \frac{\log(\frac{2}{\delta_1})}{2\epsilon_1^2} \tag{1}$$

$$m \geq \frac{2\log(\frac{2|H|}{\delta})}{\epsilon^2} \tag{2}$$

Problem 3: Naive Bayes Maximum Likelihood

(12 points)

Consider binary dataset $S \overset{i.i.d.}{\sim} D$ with observations in the form $\{(x_j^1, ..., x_j^n), y_j)\}$. Define $c(y)$ as a function that counts the number of observations such that the label is y.

$$c(y) = \sum_{(x_j, y_j) \in S} [y_j = y]$$

Define $c(i, y)$ as a function that counts the number of observations such that the label is y and $x^i = 1$.

$$c(i, y) = \sum_{(x_j, y_j) \in S} [y_j = y, x_j^i = 1]$$

Define b as $\mathbb{P}(Y = 1)$, and b^{iy} as $\mathbb{P}(X^i = 1|Y = y)$. Prove that the following estimators are MLE for these parameters:

$$\widehat{b}_{MLE} = \frac{c(1)}{|S|} \quad \text{and} \quad \widehat{b^{iy}}_{MLE} = \frac{c(i, y)}{c(y)}$$

Solution: Let $L(b, b^{iy} \mid S)$ be the likelihood of the parameters of the model.

$$\begin{aligned}
L(b, b^{iy} \mid S) &= \mathbb{P}(S \mid b, b^{iy}) \\
&= \prod_{j=1}^{n} \mathbb{P}(x_j, y_j \mid b, b^{iy}) \\
&= \prod_{j=1}^{n} \mathbb{P}(y_j \mid b, b^{iy}) \mathbb{P}(x_j \mid y_j, b, b^{iy}) \\
&= \prod_{j=1}^{n} \mathbb{P}(y_j \mid b, b^{iy}) \prod_{i=1}^{m} \mathbb{P}(x_j^i \mid y_j, b, b^{iy}) \\
&= \prod_{j=1}^{n} b^{y_j} (1 - b)^{1 - y_j} \prod_{i=1}^{m} \left(b^{iy_j}\right)^{x_j^i} \left(1 - b^{iy_j}\right)^{1 - x_j^i} \\
\log L(b, b^{iy} \mid S) &= \sum_{j=1}^{n} y_j \log(b) + (1 - y_j) \log(1 - b) + \sum_{i=1}^{m} x_j^i \log\left(b^{iy_j}\right) + \left(1 - x_j^i\right) \log\left(1 - b^{iy_j}\right)
\end{aligned}$$

Now we differentiate with respect to the different parameters and set to 0:

$$\begin{aligned}
\frac{\partial}{\partial b} \log L(b, b^{iy} \mid S) &= \frac{c(1)}{b} - \frac{c(0)}{1-b} = 0 \\
\frac{c(1)}{b} &= \frac{c(0)}{1-b} \\
\frac{c(1)}{b} &= \frac{|S| - c(1)}{1-b} \\
c(1) - bc(1) &= b|S| - bc(1) \\
c(1) &= b|S| \\
\widehat{b}_{MLE} &= \frac{c(1)}{|S|}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial b^{iy}} \log L(b, b^{iy} \mid S) &= \frac{c(i,y)}{b^{iy}} - \frac{c(y) - c(i,y)}{1-b^{iy}} = 0 \\
\frac{c(i,y)}{b^{iy}} &= \frac{c(y) - c(i,y)}{1-b^{iy}} \\
c(i,y) - b^{iy}c(i,y) &= b^{iy}c(y) - b^{iy}c(i,y) \\
c(i,y) &= b^{iy}c(y) \\
\widehat{b^{iy}}_{MLE} &= \frac{c(i,y)}{c(y)}
\end{aligned}$$

Problem 4: Gradient Descent

(18 points)

We have a convex function f over the closed interval $[-b, b]$ (for some positive number b). Let f' be the derivative of f. Let α be some positive number, which will represent a learning rate parameter.

Consider using gradient descent to find the minimum of f: We start at $x_0 = 0$. Then, at each step, we set $x_{t+1} = x_t - \alpha f'(x_t)$. If x_{t+1} falls below $-b$, we set it to $-b$, and if it goes above b, we set it to b.

We say that an optimization algorithm (such as gradient descent) *ϵ-converges* if, at some point, x_t stays within ϵ of the true minimum. Formally, we have ϵ-convergence at time t if

$$|x_{t'} - x_{\min}| \leq \epsilon, \qquad \text{where } x_{\min} = \operatorname*{argmin}_{x \in [-b,b]} f(x)$$

for all $t' \geq t$.

a. For $\alpha = 0.1$, $b = 1$, and $\epsilon = 0.001$, find a convex function f so that running gradient descent does not ϵ-converge. Specifically, make it so that $x_0 = 0$, $x_1 = b$, $x_2 = -b$, $x_3 = b$, $x_4 = -b$, etc.

b. For $\alpha = 0.1$, $b = 1$, and $\epsilon = 0.001$, find a convex function f so that gradient descent does ϵ-converge, but only after at least 10,000 steps.

c. Construct a different optimization algorithm that has the property that it will always ϵ-converge (for any convex f) within $\log_2(2b/\epsilon)$ steps.

d. Unfortunately, even if x_t is within ϵ of $x_{\min}$, $f(x_t)$ can be arbitrarily greater than $f(x_{\min})$. However, consider the case where the derivative of f is always between $-r$ and r. ($\forall x \in [-b, b]$, $f'(x) \in [-r, r]$.) In this case, we can make a guarantee about the difference between $f(x_t)$ and $f(x_{\min})$.

   Given that $|x_t - x_{\min}| \leq \epsilon$ and that $-r \leq f'(x) \leq r$, find a bound on $|f(x_t) - f(x_{\min})|$ in terms of ϵ and r.

Solution:

a. We need a function with a steep negative gradient at $x = 0$ and $x = -1$ and a steep positive gradient at $x = 1$. For example, consider

$$f(x) = \left|20x - \frac{1}{2}\right|$$

We start at $x_0 = 0$. Then we set $x_1 = x_0 - 0.1f'(x_0) = -0.1(-20) = 2$, so we set $x_1 = 1$. Then we set $x_2 = x_1 - 0.1f'(x_1) = 1 - 0.1(20) = -1$, and so on.

b. Consider $f(x) = 0.001x$. We start at $x_0 = 0$. Then we set $x_1 = x_0 - 0.1f'(x_0) = -0.1(0.001) = -0.0001$. So at each step we move 0.0001 to the left, until we reach $x_{min} = -1$ at time $t = 10000$.

c. Binary search: Start with $x_L = -b$ and $x_H = b$. At each step, set $x_t = \frac{1}{2}(x_L + x_H)$. Check $f'(x_t)$. If it's positive, then the minimum must be to the left, so set $x_H = x_t$. If it's negative, then the minimum must be to the right, so set $x_L = x_t$.

The search range $[x_L, x_H]$ has initial size $2b$ and is halved at each iteration, so after t iterations it has size $2b/2^t$. Setting this equal to ϵ (the desired search range size):

$$\begin{aligned} \epsilon &= \frac{2b}{2^t} \\ 2^t &= \frac{2b}{\epsilon} \\ t &= \log_2 \frac{2b}{\epsilon} \end{aligned}$$

So after $\log_2 (2b/\epsilon)$ iterations, we will have ϵ-convergence to the true minimum.

d. Let $a = \min(x_t, x_{min})$ and $b = \max(x_t, x_{min})$. Consider

$$f(b) - f(a) = \int_a^b f'(x)\, dx$$

If $f'(x) = r$, then this equals $(b - a)r$, which is at most ϵr. If $f'(x) = -r$, then this equals $-(b - a)r$, which is at least $-\epsilon r$. Therefore $-\epsilon r \leq f(b) - f(a) \leq \epsilon r$, so $|f(x_t) - f(x_{\min})| \leq \epsilon r$.