CS 189/289A Introduction to Machine Learning
Fall 2023 Jennifer Listgarten, Jitendra Malik

DIS7

1 The accuracy of learning decision boundaries

This problem exercises your basic probability (e.g. from 70) in the context of understanding why lots of training data helps to improve the accuracy of learning things.

For each $\theta \in (1/3, 2/3)$, define $f_\theta : [0,1] \to \{0,1\}$, such that

$$f_\theta(x) = \begin{cases} 1 \text{ if } x > \theta \\ 0 \text{ otherwise.} \end{cases}$$

The function is plotted in Figure 1.

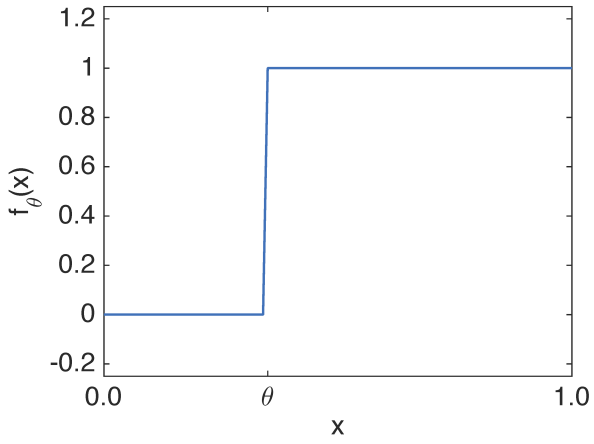

Figure 1: Plot of function $f_\theta(x)$ against x.

We draw samples $X_1, X_2, \ldots, X_n$ uniformly at random and i.i.d. from the interval $[0,1]$. Our goal is to learn an estimate for θ from n random samples $(X_1, f_\theta(X_1)), (X_2, f_\theta(X_2)), \ldots, (X_n, f_\theta(X_n))$.

Let $T_{min} = \max(\{\frac{1}{3}\} \cup \{X_i | f_\theta(X_i) = 0\})$. We know that the true θ must be larger than T_{min}.

Let $T_{max} = \min(\{\frac{2}{3}\} \cup \{X_i | f_\theta(X_i) = 1\})$. We know that the true θ must be smaller than T_{max}.

The gap between T_{min} and T_{max} represents the uncertainty we will have about the true θ given the training data that we have received.

(a) **What is the probability that $T_{max} - \theta > \epsilon$ as a function of ϵ? And what is the probability that $\theta - T_{min} > \epsilon$ as a function of ϵ?**

Solution: First note that when $\theta + \epsilon > \frac{2}{3}$ we have that $\mathbf{P}(T_{max} > \theta + \epsilon) \leq \mathbf{P}(T_{max} > \frac{2}{3}) = 0$ by definition of T_{max}. We can see this by cases: if for all X_i where $f_\theta(X_i) = 1$ we have $X_i > \frac{2}{3}$, then $T_{max} = \frac{2}{3}$; if for at least one X_i where $f_\theta(X_i) = 1$ we have $X_i < \frac{2}{3}$, then $T_{max} < \frac{2}{3}$. Hence $\mathbf{P}(T_{max}) \leq \frac{2}{3} = 1$.

For the case when $\theta+\epsilon < \frac{2}{3}$, the task is to find the probability of the event of the random variable T_{max} defined by $\mathcal{E} := \{T_{max} > \theta + \epsilon\}$. Note that because $\mathbf{P}(\{T_{max} < \theta\}) = 0$ or equivalently $\{T_{max} < \theta\} = \emptyset$, the probability $\mathbf{P}(\mathcal{E})$ can alternatively be expressed by the probability of a different event $\mathcal{E}_0 = \{\theta \leq T_{max} \leq \theta + \epsilon\}$ in terms of

$$\begin{aligned}\mathbf{P}(\mathcal{E}) &= \mathbf{P}(\{T_{max} > \theta + \epsilon\} \cup \{T_{max} < \theta\}) \\ &= 1 - \mathbf{P}(\{\theta \leq T_{max} \leq \theta + \epsilon\}) = 1 - \mathbf{P}(\mathcal{E}_0).\end{aligned}$$

Now consider the event $\mathcal{E}_1 :=$ {at least one X_i lies in $[\theta, \theta+\epsilon]$}. You can now show that $\mathcal{E}_0 = \mathcal{E}_1$, i.e.

$$\{\theta \leq T_{max} \leq \theta + \epsilon\} = \{\text{at least one } X_i \text{ lies in } [\theta, \theta + \epsilon]\}.$$

by definition of T_{max}.

Going back to the original event $\mathcal{E}$ we thus find

$$\mathbf{P}(\mathcal{E}) = 1 - \mathbf{P}(\mathcal{E}_0) = 1 - \mathbf{P}(\mathcal{E}_1) = \mathbf{P}(\mathcal{E}_1^c)$$

where the complement $\mathcal{E}_1^c = \{\text{no } X_i \text{ lies in } [\theta, \theta + \epsilon]\} = \bigcap_{i=1}^n \{X_i \notin [\theta, \theta + \epsilon]\}$.

Restating the probability of $\mathcal{E}$ in terms of an intersection of events on X_i now allows us to easily find $\mathbf{P}(\mathcal{E})$ because of independence and uniform distribution of X_i, which reads

$$\mathbf{P}(\bigcap_{i=1}^n \{X_i \notin [\theta, \theta + \epsilon]\}) = \prod_{i=1}^n \mathbf{P}(\{X_i \notin [\theta, \theta + \epsilon]\}) = (1 - \epsilon)^n.$$

In summary, we obtain

$$\mathbf{P}(T_{max} - \theta > \epsilon) = \begin{cases} (1 - \epsilon)^n & \theta + \epsilon < \frac{2}{3} \\ 0 & \text{o.w.} \end{cases}$$

Similar analysis applies to the second part, except our lower bound is $\frac{1}{3}$:

$$\mathbf{P}(\theta - T_{min} > \epsilon) = \begin{cases} (1 - \epsilon)^n & \theta - \epsilon > \frac{1}{3} \\ 0 & \text{o.w.} \end{cases}$$

(b) Suppose that you would like the estimator $\hat{\theta} = (T_{max} + T_{min})/2$ for θ that is ϵ-close (defined as $|\hat{\theta} - \theta| < \epsilon$, where $\hat{\theta}$ is the estimation and θ is the true value) with probability at least $1 - \delta$. Both ϵ and δ are some small positive numbers. **Please bound or estimate how big of an n do you need?** You do not need to find the optimal lowest sample complexity n, an approximation using results of question (a) is fine.

Solution: One way to obtain $\hat{\theta}$ within a window of size 2ϵ is to have both T_{max} and T_{min} be within ϵ of θ. To see this, define random variables $L = \theta - T_{min}, U = T_{max} - \theta$. When $L < \epsilon$ and $U < \epsilon$, we have $\theta - T_{min} < \epsilon$ and $0 < T_{max} - \theta$. Adding the two inequalities, we have $\theta - T_{min} < T_{max} - \theta + \epsilon$, thus $\hat{\theta} - \theta > -\epsilon/2 > -\epsilon$. Similarly, with those conditions, we have $\hat{\theta} - \theta < \epsilon$. Thus $L < \epsilon$ and $U < \epsilon$ is a sufficient condition for $\hat{\theta}$ to be ϵ-close, that is

$$\mathbf{P}(|\hat{\theta} - \theta| < \epsilon) \geq \mathbf{P}(\{L < \epsilon\} \cap \{U < \epsilon\}).$$

Instead of lower bounding $\mathbf{P}(\{L < \epsilon\} \cap \{U < \epsilon\})$ we upper bound the probability of the complement of the event, which reads $\{L > \epsilon\} \cup \{U > \epsilon\}$, via union bound as follows:

$$\mathbf{P}(\{L > \epsilon\} \cup \{U > \epsilon\}) \leq \mathbf{P}(\{L > \epsilon\}) + \mathbf{P}(\{U > \epsilon\}) \leq 2(1 - \epsilon)^n$$

using the result in problem (a).

We must ensure that this probability is upper bounded by δ, which ensures that we succeed with probability at least $1 - \delta$. Solving for n, we have

$$2(1 - \epsilon)^n < \delta$$
$$n > \frac{\ln\left(\frac{2}{\delta}\right)}{\ln(1/(1 - \epsilon))}.$$

Again, using the approximation $\ln(1 - x) \sim -x$, we have $n > \frac{1}{\epsilon}\ln(2/\delta)$ for ϵ small.

(c) Let us say that instead of getting random samples $(X_i, f(X_i))$, we were allowed to choose where to sample the function, but you had to choose all the places you were going to sample in advance. **Propose a method to estimate θ. How many samples suffice to achieve an estimate that is ϵ-close as above?** (**Hint:** You need not use a randomized strategy.)

Solution: Pick n points uniformly spaced on the interval $(\frac{1}{3}, \frac{2}{3})$. Then, the ith sample $X_i = \frac{1}{3} + \frac{i}{3n}$. Since we have n points, we create intervals of length $\frac{1}{3n}$. If our intervals are smaller than 2ϵ, we can guarantee that we estimate θ within an interval of 2ϵ. Solving for n, we have $\frac{1}{3n} < 2\epsilon$ and so $n > \frac{1}{6\epsilon}$ samples are sufficient.

Note that using our calculations the sample complexity for this deterministic method is *always lower* than the sample complexity of the probabilistic method in problem (b) $\delta < 1$ since $\ln\left(\frac{2}{\delta}\right) > \frac{1}{6}$ for any $\delta < 1$. Therefore, uniform sampling and allowing for some non-zero probability that we do not obtain an ϵ-close estimator, does not require fewer samples than a deterministic method which always ensures an ϵ-close estimator. In many other settings however, allowing some uncertainty (of finding a good estimator) can help to reduce the sample complexity significantly.

(d) Suppose that you could pick where to sample the function adaptively — choosing where to sample the function in response to what the answers were previously. **Propose a method to estimate θ. How many samples suffice to achieve an estimate that is ϵ-close as above?**

Solution: Use binary search: start with three pointers, $s = 1/3, e = 2/3$ with m as the midpoint. If $f(m) = 0$, set $s = m$ and recompute the midpoint (i.e., search over the second half of the range). Otherwise, $f(m) = 1$ and set $e = m$ (i.e., search over the first half of the range). For each point sampled, we reduce the size of the range by half, so after n points, the interval we consider is $\frac{1}{3 \cdot 2^n}$. We want this to be less than 2ϵ, and so

$$\frac{1}{3 \cdot 2^n} < 2\epsilon \implies n > \log_2(\frac{1}{3\epsilon}) - 1.$$

(e) In the three sampling approaches above: random, deterministic, and adaptive, **compare the scaling of n with ϵ (and δ as well for the random case).**

Solution:

(a) For random, n is logarithmic in $1/\delta$. For ϵ, we use the approximation $\ln(1/(1-\epsilon)) \sim \epsilon$ to conclude that n is inversely related to ϵ.

(b) For deterministic, n is inversely related to ϵ. Note that this is the same scaling as choosing random evaluation points.

(c) For adaptive, n is logarithmic in $\frac{1}{\epsilon}$.

(f) **Why do you think we asked this series of questions? What are the implications of those results in a machine learning application?**

Solution: We ask this question because we want to show how the number of training examples affects the accuracy. Intuitively, more data lead to a more accurate estimator. We quantify this intuition with a simple but concrete example.

The three sampling approaches are some common ways to get the training data. When most of the real world datasets are collected, one doesn't have any control on X_i. That is the random sampling paradigm. The deterministic sampling paradigm refers to the scenario when one could carefully design a set of X_i. One might think that the sample complexity of the deterministic case should be much better than that of the random one, however for this particular model, they are not quite different. There is only a factor of $\log \frac{1}{\delta}$ off. However, when we move to the adaptive paradigm, the sample complexity is exponentially smaller.

For practical machine learning applications, the implication is that you want to have as much control on the samples as you can (such as adaptive sampling) to learn a better model with the same amount of data.

2 The Classical Bias-Variance Tradeoff

Consider a random variable X, which has unknown mean μ and unknown variance σ^2. Given n iid realizations of training samples $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ from the random variable, we wish to estimate the mean of X. We will call our estimate of μ the random variable $\hat{X}$, which has mean $\hat{\mu}$. There are a few ways we can estimate μ given the realizations of the n samples:

1. Average the n samples: $\frac{x_1+x_2+\ldots+x_n}{n}$.
2. Average the n samples and one sample of 0: $\frac{x_1+x_2+\ldots+x_n}{n+1}$.
3. Average the n samples and n_0 samples of 0: $\frac{x_1+x_2+\ldots+x_n}{n+n_0}$.
4. Ignore the samples: just return 0.

In the parts of this question, we will measure the *bias* and *variance* of each of our estimators. The *bias* is defined as

$$\mathbb{E}[\hat{X} - \mu]$$

and the *variance* is defined as

$$\text{Var}[\hat{X}].$$

(a) What is the bias of each of the four estimators above?

Solution: $\mathbb{E}[\hat{X} - \mu] = \mathbb{E}[\hat{X}] - \mu$, so we have the following biases:

(a) $\mathbb{E}[\hat{X}] = \mathbb{E}[\frac{X_1+X_2+\ldots+X_n}{n}] = \frac{n\mu}{n} \implies \text{bias} = 0$
(b) $\mathbb{E}[\hat{X}] = \mathbb{E}[\frac{X_1+X_2+\ldots+X_n}{n+1}] = \frac{n\mu}{n+1} \implies \text{bias} = -\frac{1}{n+1}\mu$
(c) $\mathbb{E}[\hat{X}] = \mathbb{E}[\frac{X_1+X_2+\ldots+X_n}{n+n_0}] = \frac{n\mu}{n+n_0} \implies \text{bias} = -\frac{n_0}{n+n_0}\mu$
(d) $\mathbb{E}[\hat{X}] = 0 \implies \text{bias} = -\mu$

(b) What is the variance of each of the four estimators above?

Solution: The two key identities to remember are $\text{Var}[A + B] = \text{Var}[A] + \text{Var}[B]$ (when A and B are independent) and $\text{Var}[kA] = k^2\,\text{Var}[A]$, where A and B are random variables and k is a constant.

(a) $\text{Var}[\hat{X}] = \text{Var}[\frac{X_1+X_2+\ldots+X_n}{n}] = \frac{1}{n^2}\,\text{Var}[X_1 + X_2 + \ldots + X_n] = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}$
(b) $\text{Var}[\hat{X}] = \text{Var}[\frac{X_1+X_2+\ldots+X_n}{n+1}] = \frac{1}{(n+1)^2}\,\text{Var}[X_1 + X_2 + \ldots + X_n] = \frac{1}{(n+1)^2}(n\sigma^2) = \frac{n}{(n+1)^2}\sigma^2$
(c) $\text{Var}[\hat{X}] = \text{Var}[\frac{X_1+X_2+\ldots+X_n}{n+n_0}] = \frac{1}{(n+n_0)^2}\,\text{Var}[X_1 + X_2 + \ldots + X_n] = \frac{1}{(n+n_0)^2}(n\sigma^2) = \frac{n}{(n+n_0)^2}\sigma^2$
(d) $\text{Var}[\hat{X}] = 0$

(c) Suppose we have constructed an estimator $\hat{X}$ from some samples of X. We now want to know how well $\hat{X}$ estimates a new independent sample of X. Denote this new sample by X'. Derive a general expression for $\mathbb{E}[(\hat{X} - X')^2]$ in terms of σ^2 and the bias and variance of the estimator

$\hat{X}$. Similarly, derive an expression for $\mathbb{E}[(\hat{X}-\mu)^2]$. Compare the two expressions and comment on the differences between them.

Solution: Since $\hat{X}$ is a function of X, we conclude that the random variables $\hat{X}$ and X' are independent of each other. Now we provide two ways to solve the first problem.

Method 1: In this method, we use the trick of adding and subtracting a term to derive the desired expression:

$$\begin{aligned}
\mathbb{E}[(\hat{X}-X')^2] &= \mathbb{E}[(\hat{X}-\mu+\mu-X')^2] \\
&= \mathbb{E}[(\hat{X}-\mu)^2] + \underbrace{E[(\mu-X')^2]}_{=\text{Var}(X')=\sigma^2} \\
&= \mathbb{E}[(\hat{X}-\mu)^2] + \sigma^2 \\
&= \mathbb{E}[(\hat{X}-E[\hat{X}]+E[\hat{X}]-\mu)^2] + \sigma^2 \\
&= \underbrace{\mathbb{E}[(\hat{X}-\mathbb{E}[\hat{X}])^2]}_{=\text{Var}(\hat{X})} + \underbrace{(\mathbb{E}[\hat{X}]-\mu)^2}_{=\text{bias}^2} + 2\underbrace{\mathbb{E}[(\hat{X}-E[\hat{X}])\cdot(\mathbb{E}[\hat{X}]-\mu)]}_{=0} + \sigma^2
\end{aligned}$$

Method 2: In this method, we make use of the definition of variance. We have

$$\begin{aligned}
\mathbb{E}[(\hat{X}-X')^2] &= \mathbb{E}[\hat{X}^2] + \mathbb{E}[X'^2] - 2\,\mathbb{E}[\hat{X}X'] \\
&= (\text{Var}(\hat{X}) + (E[\hat{X}])^2) + (\text{Var}(X') + (\mathbb{E}[X'])^2) - 2\underbrace{\mathbb{E}[\hat{X}]\,\mathbb{E}[X']}_{\text{independence}} \\
&= (\mathbb{E}[\hat{X}]^2 - 2\,\mathbb{E}[\hat{X}]\,\mathbb{E}[X'] + \mathbb{E}[X']^2) + \text{Var}(\hat{X}) + \underbrace{\text{Var}(X')}_{=\text{Var}(X)} \\
&= (\mathbb{E}[\hat{X}] - \underbrace{\mathbb{E}[X']}_{=\mathbb{E}[X]=\mu})^2 + \text{Var}(\hat{X}) + \text{Var}(X) \\
&= \underbrace{(\mathbb{E}[\hat{X}]-\mu)^2}_{=\text{bias}^2} + \text{Var}(\hat{X}) + \sigma^2
\end{aligned}$$

The first term is equivalent to the bias of our estimator squared, the second term is the variance of the estimator, and the last term is the irreducible error.

Now let's do $\mathbb{E}[(\hat{X}-\mu)^2]$.

$$\begin{aligned}
\mathbb{E}[(\hat{X}-\mu)^2] &= \mathbb{E}[\hat{X}^2] + \mathbb{E}[\mu^2] - 2\,\mathbb{E}[\hat{X}\mu] && (1)\\
&= (\text{Var}(\hat{X}) + \mathbb{E}[\hat{X}]^2) + (\text{Var}(\mu) + \mathbb{E}[\mu]^2) - 2\,\mathbb{E}[\hat{X}\mu] && (2)\\
&= (\mathbb{E}[\hat{X}]^2 - 2\,\mathbb{E}[\hat{X}\mu] + E[\mu]^2) + \text{Var}(\hat{X}) + \text{Var}(\mu) && (3)\\
&= (\mathbb{E}[\hat{X}] - \mathbb{E}[\mu])^2 + \text{Var}(\hat{X}) + \text{Var}(\mu) && (4)\\
&= (\mathbb{E}[\hat{X}]-\mu)^2 + \text{Var}(\hat{X}). && (5)
\end{aligned}$$

Notice that these two expected squared errors resulted in the same expressions except for the σ^2 in $\mathbb{E}[(\hat{X}-X')^2]$. The error σ^2 is considered "irreducible error" because it is associated with

the noise that comes from sampling from the distribution of X. This term is not present in the second derivation because μ is a fixed value that we are trying to estimate.

(d) It is a common mistake to assume that an unbiased estimator is always "best." Let's explore this a bit further. Compute $E[(\hat{X} - \mu)^2]$ for each of the estimators above. **Solution:** Adding the previous two answers:

(a) $\frac{\sigma^2}{n}$

(b) $\frac{1}{(n+1)^2}(\mu^2 + n\sigma^2)$

(c) $\frac{1}{(n+n_0)^2}(n_0^2\mu^2 + n\sigma^2)$

(d) μ^2

(e) Demonstrate that the four estimators are each just special cases of the third estimator, but with different instantiations of the hyperparameter n_0.

Solution: The derivation for the third estimator works for *any* value of n_0. The first estimator is just the third estimator with n_0 set to 0:

$$\frac{x_1 + x_2 + \ldots + x_n}{n + n_0} = \frac{x_1 + x_2 + \ldots + x_n}{n + 0} + \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

The second estimator is just the third estimator with n_0 set to 1:

$$\frac{x_1 + x_2 + \ldots + x_n}{n + n_0} = \frac{x_1 + x_2 + \ldots + x_n}{n + 1}.$$

The last estimator is the limiting behavior as n_0 goes to ∞. In other words, we can get arbitrarily close to the fourth estimator by setting n_0 very large:

$$\lim_{n_0 \to \infty} \frac{x_1 + x_2 + \ldots + x_n}{n + n_0} = 0.$$

(f) What happens to bias as n_0 increases? What happens to variance as n_0 increases?

Solution:

One reason for increasing the samples of n_0 is if you have reason to believe that X is centered around 0. In increasing the number of zeros we are injecting more confidence in our belief that the distribution is centered around zero. Consequently, in increasing the number of "fake" data, the variance decreases because your distribution becomes more peaked. Examining the expressions for bias and variance for the third estimator, we can see that larger values of n_0 result in decreasing variance ($\frac{n}{(n+n_0)^2}\sigma^2$) but potentially increasing bias ($\frac{n_0\mu}{n+n_0}$). Hopefully you can see that there is a trade-off between bias and variance. Using an unbiased estimator is not always optimal nor is using an estimator with small variance always optimal. One has to carefully trade-off the two terms in order to obtain minimum squared error.