CS 189/289A Introduction to Machine Learning
Fall 2024 Jennifer Listgarten, Saeed Saremi

DIS8

1 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 $\mathbb{E}[(\hat{X} - \mu)^2]$.

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

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

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.

2 Decision Trees

Consider constructing a decision tree on data with d features and n training points where each feature is real-valued and each label takes one of m possible values. The splits are two-way, and are chosen to maximize the information gain. We only consider splits that form a linear boundary parallel to one of the axes. We will only consider a standalone decision tree and not a random forest (hence no randomization). Recall the definition of information gain:

$$IG(\textbf{node}) = H(S) - \frac{|S_l|H(S_l) + |S_r|H(S_r)}{|S_l| + |S_r|}$$

where S is set of samples considered at **node**, S_l is the set of samples remaining in the left sub-tree after **node**, S_r is the set of samples remaining in the right sub-tree after **node**, and $H(S)$ is the entropy over a set of samples:

$$H(S) = -\sum_{i=1}^{C} p_i \log(p_i)$$

Here, C is the number of classes, and p_i is the proportion of samples in S labeled as class i.

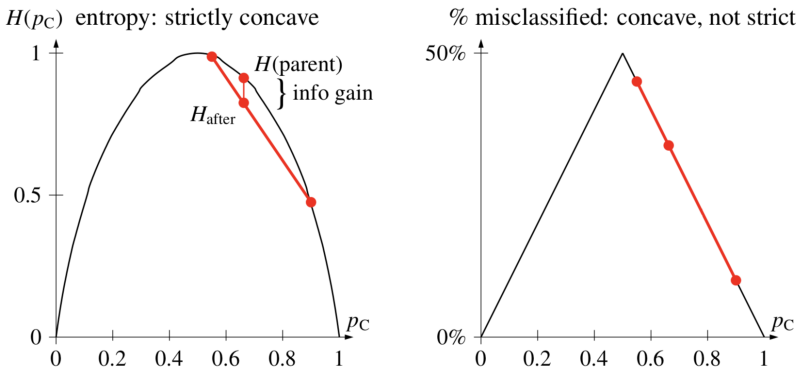
(a) Intuitively, how does the bias-variance trade-off relate to the depth of a decision tree?

Solution: If a decision tree is very deep, the model is likely to overfit, and thereby increase variance. Intuitively, there are many conditions checked before making a decision, which makes the decision rule too fine-grained and sensitive to small perturbations; for example, if only one of the many conditions is not satisfied, this might result in a completely different prediction. On the other hand, if the tree is very shallow, this might increase bias. In this case, the decisions are too "coarse".

(b) Draw the graph of entropy $H(p_c)$ when there are only two classes C and D, with $p_D = 1 - p_C$. Is the entropy function strictly concave, concave, strictly convex, or convex? Why? What is the significance?

Hint: For the significance, recall the information gain.

Solution: The function is strictly concave. Notice that the function $-x \log x$ is strictly concave



in $[0, 1]$, and a sum of strictly concave functions is strictly concave.

Significance: Suppose we pick two points on the entropy curve, then draw a line segment connecting them. Because the entropy curve is strictly concave, the interior of the line segment is strictly below the curve. Any point on that segment represents a weighted average of the two entropies for suitable weights. If you unite the two sets into one parent set, the parent set's value p_C is the weighted average of the children's p_c's. Therefore, the point directly above that point on the curve represents the parent's entropy. The information gain is the vertical distance between them. So the information gain is positive unless the two child sets both have exactly the same p_C and lie at the same point on the curve. Note that this is why we can also pick even simpler strictly concave functions (like $H'(p) = p(1-p)$) which will work nearly as well.

On the other hand, for the graph on the right, plotting the % misclassified, if we draw a line segment connecting two points on the curve, the segment might lie entirely on the curve. In that case, uniting the two child sets into one, or splitting the parent set into two, changes neither the total misclassified sample points nor the weighted average of the % misclassified. The bigger problem, though, is that many different splits will get the same weighted average cost; this test doesn't distinguish the quality of different splits well.

(c) Prove or give a counter-example: The information gain at the root is at least as much as the information gain at any other node.
Hint: Think about the XOR function. Precisely, consider the set

$$S = \{(0,0;0),(0,1;1),(1,0;1),(1,1;0)\},$$

where the first two entries in every sample are features, and the last one is the label.

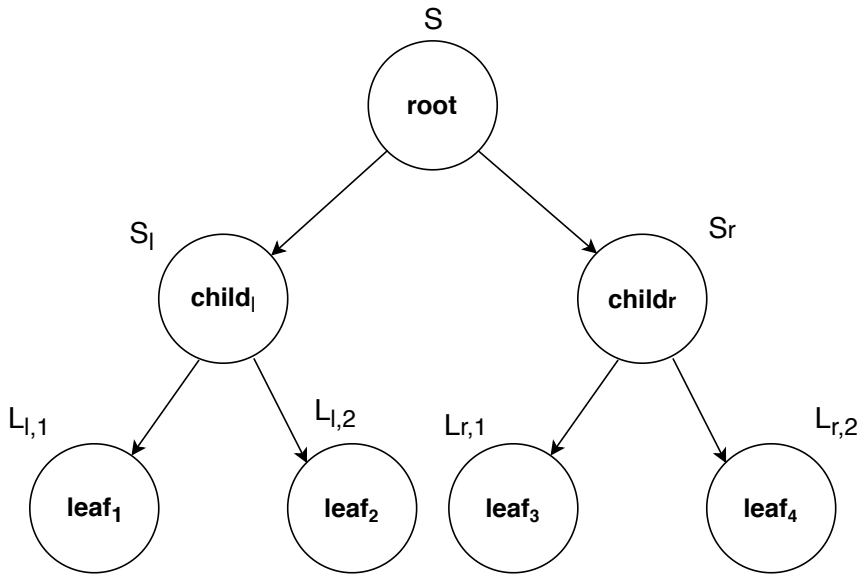
Solution: False. As suggested in the hint, we consider the XOR function. Then, $H(S) = 1$. The first split is done based on the first feature, which gives $S_l = \{(0,0;0),(0,1;1)\}$ and $S_r = \{(1,0;1),(1,1;0)\}$; denote the corresponding nodes as $\textbf{child}_l$ and $\textbf{child}_r$ respectively. This gives $H(S_l) = 1$ and $H(S_r) = 1$. The information gain of the first split is:

$$IG(\textbf{root}) = H(S) - \frac{|S_l|H(S_l) + |S_r|H(S_r)}{|S_l| + |S_r|} = 0$$

Now we further split S_l and S_r according to the second feature, which gives 4 leaves of 1 sample each. Denote the leaf samples corresponding to S_r as $L_{r,1}$ and $L_{r,2}$, and accordingly denote by $L_{l,1}$ and $L_{l,2}$ the leaves corresponding to S_l. Now we have

$$IG(\textbf{child}_l) = H(S_l) - \frac{1 \cdot H(L_{l,1}) + 1 \cdot H(L_{l,2})}{1+1} = 1$$

and analogously $IG(\textbf{child}_r) = 1$. Therefore, the information gain at each of the child nodes is 1, while the information gain at the root is 0.



(d) Prove or give a counter-example: In any path from the root to a leaf, the same feature will never be split on twice.

Solution: False. Example: one dimensional feature space with training points of two classes x and o arranged as xxxooooxxx.

3 Curse of Dimensionality in Nearest Neighbor Classification

We have a training set: $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})$, where $\mathbf{x}^{(i)} \in \mathbb{R}^d$. To classify a new point $\mathbf{x}$, we can use the nearest neighbor classifier:

$$\text{class}(\mathbf{x}) = y^{(i^*)} \quad \text{where } \mathbf{x}^{(i^*)} \text{ is the nearest neighbor of } \mathbf{x}.$$

Assume any data point $\mathbf{x}$ that we may pick to classify is inside the Euclidean ball of radius 1, i.e. $\|\mathbf{x}\|_2 \leq 1$. To be confident in our prediction, in addition to choosing the class of the nearest neighbor, we want the distance between $\mathbf{x}$ and its nearest neighbor to be small, within some positive ϵ:

$$\|\mathbf{x} - \mathbf{x}^{(i^*)}\|_2 \leq \epsilon \quad \text{for all } \|\mathbf{x}\|_2 \leq 1. \tag{6}$$

What is the minimum number of training points we need for inequality (6) to hold (assuming the training points are well spread)? How does this lower bound depend on the dimension d?

Hint: Think about the volumes of the hyperspheres in d dimensions.

Solution: Let B_0 be the ball centered at the origin, having radius 1 (inside which we assume our data lies). Let $B_i(\epsilon)$ be the ball centered at $\mathbf{x}^{(i)}$, having radius ϵ. For inequality (6) to hold, for any point $\mathbf{x} \in B_0$, there must be at least one index i such that $\mathbf{x} \in B_i(\epsilon)$. This is equivalent to saying that the union of $B_1(\epsilon), \ldots, B_n(\epsilon)$ covers the ball B_0. Let $\text{vol}(B)$ indicate the volume of object B, then we have

$$\sum_{i=1}^{n} \text{vol}(B_i(\epsilon)) = n\text{vol}(B_1(\epsilon)) \geq \text{vol}(\cup_{i=1}^{n} B_i(\epsilon)) \geq \text{vol}(B_0).$$

where the last inequality holds because we are assuming the union of $B_1(\epsilon), \ldots, B_n(\epsilon)$ covers the ball B_0. This implies

$$n \geq \frac{\text{vol}(B_0)}{\text{vol}(B_1(\epsilon))} = \frac{c(1^d)}{c\epsilon^d} = \frac{1}{\epsilon^d}$$

Where the constant c is dependent on the formula for the volume of a hypersphere in d dimensions.

Note that we can pick $\frac{1}{\epsilon^d}$ training points and still satisfy (6) only if all the training points are well spread (the union of $B_1(\epsilon), \ldots, B_n(\epsilon)$ covers the ball B_0).

This lower bound suggests that to make an accurate prediction on high-dimensional input, we need exponentially many samples in the training set. This exponential dependence is sometimes called the *curse of dimensionality*. It highlights the difficulty of using non-parametric methods for solving high-dimensional problems.