CS 189 Introduction to Machine Learning
Spring 2022 Marvin Zhang

DIS7

1 Risk Minimization with Doubt

Suppose we have a classification problem with classes labeled $1, \dots, c$ and an additional "doubt" category labeled $c+1$. Let $f : \mathbb{R}^d \to \{1, \dots, c+1\}$ be a decision rule. Define the loss function

$$L(f(\mathbf{x}), y) = \begin{cases} 0 & \text{if } f(\mathbf{x}) = y \quad f(\mathbf{x}) \in \{1, \dots, c\}, \\ \lambda_c & \text{if } f(\mathbf{x}) \neq y \quad f(\mathbf{x}) \in \{1, \dots, c\}, \\ \lambda_d & \text{if } f(\mathbf{x}) = c+1 \end{cases} \tag{1}$$

where $\lambda_c \geq 0$ is the loss incurred for making a misclassification and $\lambda_d \geq 0$ is the loss incurred for choosing doubt. In words this means the following:

- When you are correct, you should incur no loss.
- When you are incorrect, you should incur some penalty λ_c for making the wrong choice.
- When you are unsure about what to choose, you might want to select a category corresponding to "doubt" and you should incur a penalty λ_d.

The risk of classifying a new data point $\mathbf{x}$ as class $f(\mathbf{x}) \in \{1, 2, \dots, c+1\}$ is

$$R(f(\mathbf{x})|\mathbf{x}) = \sum_{i=1}^{c} L(f(\mathbf{x}), i)\, P(Y = i|\mathbf{x}).$$

(a) Show that the following policy $f_{opt}(x)$ obtains the minimum risk:

- (**R1**) Find class i such that $P(Y = i|\mathbf{x}) \geq P(Y = j|\mathbf{x})$ for all j, meaning you pick the class with the highest probability given x.
- (**R2**) Choose class i if $P(Y = i|\mathbf{x}) \geq 1 - \frac{\lambda_d}{\lambda_c}$
- (**R3**) Choose doubt otherwise.

Solution:

- Let's first simplify the risk given our specific loss function. If $f(\mathbf{x}) = i$ where i is not doubt, then the risk is

$$R(f(\mathbf{x}) = i|\mathbf{x}) = \sum_{j=1}^{c} L(f(\mathbf{x}) = i, y = j) P(Y = j|\mathbf{x}) \tag{2}$$

$$= 0 \cdot P(Y = i|\mathbf{x}) + \lambda_c \sum_{j=1, j\neq i} P(Y = j|\mathbf{x}) \tag{3}$$

$$= \lambda_c \left(1 - P(Y = i|\mathbf{x})\right) \tag{4}$$

When $f(\mathbf{x}) = c + 1$, meaning you've chosen doubt, the risk is:

$$R(f(\mathbf{x}) = c + 1|\mathbf{x}) = \sum_{j=1}^{c} L(f(\mathbf{x}) = c + 1, y = j)P(Y = j|\mathbf{x}) \tag{5}$$

$$= \lambda_d \sum_{j=1} P(Y = j|\mathbf{x}) \tag{6}$$

$$= \lambda_d \tag{7}$$

because $\sum_{j=1} P(Y = j|\mathbf{x})$ should sum to 1 since its a proper probability distribution.
Now let $f_{opt} : \mathbb{R}^d \to \{1, \ldots, c + 1\}$ be the decision rule which implements (**R1**)–(**R3**). We want to show that in expectation the rule f_{opt} is at least as good as an arbitrary rule f. Let $\mathbf{x} \in \mathbb{R}^d$ be a data point, which we want to classify. Let's examine all the possible scenarios where $f_{opt}(\mathbf{x})$ and another arbitrary rule $f(\mathbf{x})$ might differ:
Case 1: Let $f_{opt}(\mathbf{x}) = i$ where $i \neq c + 1$.

- Case 1a: $f(\mathbf{x}) = k$ where $k \neq i$. Then we get with (**R1**) that

$$\begin{aligned} R(f_{opt}(\mathbf{x}) = i|\mathbf{x}) &= \lambda_c \left(1 - P(Y = i|\mathbf{x})\right) \\ &\leq \lambda_c \left(1 - P(Y = k|\mathbf{x})\right) = R(f(\mathbf{x}) = k|\mathbf{x}). \end{aligned}$$

- Case 1b: $f(\mathbf{x}) = c + 1$. Then we get with (**R1**) that

$$\begin{aligned} R(f_{opt}(\mathbf{x}) = i|\mathbf{x}) &= \lambda_c \left(1 - P(Y = i|\mathbf{x})\right) \\ &\leq \lambda_c(1 - (1 - \frac{\lambda_d}{\lambda_c})) = \lambda_d = R(f(\mathbf{x}) = c + 1|\mathbf{x}). \end{aligned}$$

Case 2: Let $f_{opt}(\mathbf{x}) = c + 1$ and $f(\mathbf{x}) = k$ where $k \neq c + 1$. Then:

$$R(f(\mathbf{x}) = k|\mathbf{x}) = \lambda_c(1 - P(Y = k|\mathbf{x}))$$

$$R(f_{opt}(\mathbf{x}) = c + 1|\mathbf{x}) = \lambda_d$$

We are in case (**R3**) which means that:

$$\max_{j\in\{1,\ldots,c\}} P(Y = j|\mathbf{x}) < 1 - \lambda_d/\lambda_c$$

hence $P(Y = k|\mathbf{x}) < 1 - \lambda_d/\lambda_c$, which means

$$R(f(\mathbf{x}) = k|\mathbf{x}) > \lambda_d = R(f_{opt}(\mathbf{x}))$$

Therefore in every case we proved that the rule f_{opt} is at least as good as the arbitrary rule f, which proves that f_{opt} is an optimal rule.

(b) How would you modify your optimum decision rule if $\lambda_d = 0$? What happens if $\lambda_d > \lambda_c$? Explain why this is or is not consistent with what one would expect intuitively.

Solution: If $\lambda_d = 0$, then property (1) will hold iff there exists an $i \in \{1, \ldots, c\}$ such that $P(f_{opt}(\mathbf{x}) = i|\mathbf{x}) = 1$. So we will either classify x in class i if we are 100% sure about this, or else we will choose doubt. Of course this is completely consistent with our intuition, because choosing doubt does not have any penalty at all, since $\lambda_d = 0$.

If $\lambda_d > \lambda_c$, then we will always classify x in the class $i \in \{1, \ldots, c\}$ which gives the highest probability of correct classification. Once again this makes sense, since the cost of choosing doubt is higher than classifying $\mathbf{x}$ in any of the classes, hence our best option is to classify x in the class which gives the highest probability for a correct classification.

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

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

and the *variance* is defined as

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

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

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

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

(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 $E[(\hat{X} - X')^2]$ in terms of σ^2 and the bias and variance of the estimator

$\hat{X}$. Similarly, derive an expression for $E[(\hat{X}-\mu)^2]$. 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:

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

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

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

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

$$\begin{aligned}
E[(\hat{X}-\mu)^2] &= E[\hat{X}^2] + E[\mu^2] - 2E[\hat{X}\mu] && (8)\\
&= (Var(\hat{X}) + E[\hat{X}]^2) + (Var(\mu) + E[\mu]^2) - 2E[\hat{X}\mu] && (9)\\
&= (E[\hat{X}]^2 - 2E[\hat{X}\mu] + E[\mu]^2) + Var(\hat{X}) + Var(\mu) && (10)\\
&= (E[\hat{X}] - E[\mu])^2 + Var(\hat{X}) + Var(\mu) && (11)\\
&= (E[\hat{X}]-\mu)^2 + Var(\hat{X}). && (12)
\end{aligned}$$

Notice that these two expected squared errors resulted in the same expressions except for the σ^2 in $E[(\hat{X}-X')^2]$. 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 distriubtion 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.