

Week 1 Lecture: Concept Check Exercises

Starred problems are optional.

Statistical Learning Theory

1. Suppose $\mathcal{A} = \mathcal{Y} = \mathbb{R}$ and \mathcal{X} is some other set. Furthermore, assume $P_{\mathcal{X} \times \mathcal{Y}}$ is a discrete joint distribution. Compute a Bayes decision function when the loss function $\ell : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}$ is given by

$$\ell(a, y) = \mathbf{1}(a \neq y),$$

the 0 – 1 loss.

2. (★) Suppose $\mathcal{A} = \mathcal{Y} = \mathbb{R}$, \mathcal{X} is some other set, and $\ell : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}$ is given by $\ell(a, y) = (a - y)^2$, the square error loss. What is the Bayes risk and how does it compare with the variance of Y ?
3. Let $\mathcal{X} = \{1, \dots, 10\}$, let $\mathcal{Y} = \{1, \dots, 10\}$, and let $\mathcal{A} = \mathcal{Y}$. Suppose the data generating distribution, P , has marginal $X \sim \text{Unif}\{1, \dots, 10\}$ and conditional distribution $Y|X = x \sim \text{Unif}\{1, \dots, x\}$. For each loss function below give a Bayes decision function.

(a) $\ell(a, y) = (a - y)^2$,

(b) $\ell(a, y) = |a - y|$,

(c) $\ell(a, y) = \mathbf{1}(a \neq y)$.

4. Show that the empirical risk is an unbiased and consistent estimator of the Bayes risk. You may assume the Bayes risk is finite.
5. Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} = \mathcal{A} = \mathbb{R}$. Suppose you receive the (x, y) data points $(0, 5)$, $(.2, 3)$, $(.37, 4.2)$, $(.9, 3)$, $(1, 5)$. Throughout assume we are using the 0 – 1 loss.
 - (a) Suppose we restrict our decision functions to the hypothesis space \mathcal{F}_1 of constant functions. Give a decision function that minimizes the empirical risk over \mathcal{F}_1 and the corresponding empirical risk. Is the empirical risk minimizing function unique?
 - (b) Suppose we restrict our decision functions to the hypothesis space \mathcal{F}_2 of piecewise-constant functions with at most 1 discontinuity. Give a decision function that minimizes the empirical risk over \mathcal{F}_2 and the corresponding empirical risk. Is the empirical risk minimizing function unique?

6. (★) Let $\mathcal{X} = [-10, 10]$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$ and suppose the data generating distribution has marginal distribution $X \sim \text{Unif}[-10, 10]$ and conditional distribution $Y|X = x \sim \mathcal{N}(a + bx, 1)$ for some fixed $a, b \in \mathbb{R}$. Suppose you are also given the following data points: $(0, 1)$, $(0, 2)$, $(1, 3)$, $(2.5, 3.1)$, $(-4, -2.1)$.
- Assuming the 0 – 1 loss, what is the Bayes risk?
 - Assuming the square error loss $\ell(a, y) = (a - y)^2$, what is the Bayes risk?
 - Using the full hypothesis space of all (measurable) functions, what is the minimum achievable empirical risk for the square error loss.
 - Using the hypothesis space of all affine functions (i.e., of the form $f(x) = cx + d$ for some $c, d \in \mathbb{R}$), what is the minimum achievable empirical risk for the square error loss.
 - Using the hypothesis space of all quadratic functions (i.e., of the form $f(x) = cx^2 + dx + e$ for some $c, d, e \in \mathbb{R}$), what is the minimum achievable empirical risk for the square error loss.

Stochastic Gradient Descent

- When performing mini-batch gradient descent, we often randomly choose the mini-batch from the full training set without replacement. Show that the resulting mini-batch gradient is an unbiased estimate of the gradient of the full training set. Here we assume each decision function f_w in our hypothesis space is determined by a parameter vector $w \in \mathbb{R}^d$.
- You want to estimate the average age of the people visiting your website. Over a fixed week we will receive a total of N visitors (which we will call our full population). Suppose the population mean μ is unknown but the variance σ^2 is known. Since we don't want to bother every visitor, we will ask a small sample what their ages are. How many visitors must we randomly sample so that our estimator $\hat{\mu}$ has variance at most $\epsilon > 0$?
- (★) Suppose you have been successfully running mini-batch gradient descent with a full training set size of 10^5 and a mini-batch size of 100. After receiving more data your full training set size increases to 10^9 . Give a heuristic argument as to why the mini-batch size need not increase even though we have 10000 times more data.