

NYU Center for Data Science: DS-GA 1003

Machine Learning and Computational Statistics (Spring 2019)

Brett Bernstein

April 2, 2019

Instructions: Following most lab and lecture sections, we will be providing concept checks for review. Each concept check will:

- List the lab/lecture learning objectives. You will be responsible for mastering these objectives, and demonstrating mastery through homework assignments, exams (midterm and final), and on the final course project.
- Include concept check questions. These questions are intended to reinforce the lab/lectures, and help you master the learning objectives.

You are strongly encourage to complete all concept check questions, and to discuss these (and related) problems on Piazza and at office hours. However, problems marked with a (★) are considered optional.

Lecture 1: Introduction to Statistical Learning Theory

Topic 1: Statistical Learning Theory

Learning Objectives

1. Identify the input, action, and outcome spaces for a given machine learning problem.
2. Provide an example for which the action space and outcome spaces are the same and one for which they are different.
3. Explain the relationships between the decision function, the loss function, the input space, the action space, and the outcome space.
4. Define the risk of a decision function and a Bayes decision function.
5. Provide example decision problems for which the Bayes risk is 0 and the Bayes risk is nonzero.
6. Know the Bayes decision functions for square loss and multiclass 0/1 loss.

7. Define the empirical risk for a decision function and the empirical risk minimizer.
8. Explain what a hypothesis space is, and how it can be used with constrained empirical risk minimization to control overfitting.

Concept Check Questions

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: $(0, 1)$, $(0, 2)$, $(1, 3)$, $(2.5, 3.1)$, $(-4, -2.1)$.

- (a) Assuming the 0 – 1 loss, what is the Bayes risk?
- (b) Assuming the square error loss $\ell(a, y) = (a - y)^2$, what is the Bayes risk?
- (c) Using the full hypothesis space of all (measurable) functions, what is the minimum achievable empirical risk for the square error loss?
- (d) 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?
- (e) 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?