1 Introduction to Statistical Learning Theory

1.1 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.

1.2 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 $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ is given by

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

   the $0-1$ loss.

2. ($\star$) Suppose $\mathcal{A} = \mathcal{Y} = \mathbb{R}$, $\mathcal{X}$ is some other set, and $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ 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 $\mathcal{X} = \{1, \ldots, 10\}$, let $\mathcal{Y} = \{1, \ldots, 10\}$, and let $A = \mathcal{Y}$. Suppose the data generating distribution, P, has marginal $X \sim \text{Unif}\{1, \ldots, 10\}$ and conditional distribution $Y|X = x \sim \text{Unif}\{1, \ldots, x\}$. 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. ($\star$) 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?

2 Excess Risk Decomposition, Stochastic Gradient Descent and Regularization

2.1 Topic 1: Excess Risk Decomposition

2.1.1 Learning Objectives

1. Give precise definitions for excess risk, approximation error, estimation error, and optimization error.

2. Suppose we have nested hypothesis spaces, say $\mathcal{H}_1 \subset \mathcal{H}_2$. Explain how we would expect the approximation error and estimation error to change when we change from $\mathcal{H}_1$ to $\mathcal{H}_2$, all else fixed.

3. Explain how we would expect the approximation error and estimation error to change when we increase the sample size, all else fixed.

4. Explain optimization error, and write down an excess risk decomposition that incorporates approximation error, estimation error, and optimization error. Why might we have negative optimization error but never negative estimation error?

2.1.2 Concept Check Questions

1. Let $\mathcal{X} = \mathcal{Y} = \{1, 2, \ldots, 10\}$, $\mathcal{A} = \{1, \ldots, 10, 11\}$ and suppose the data distribution has marginal distribution $X \sim \text{Unif}\{1, \ldots, 10\}$. Furthermore, assume $Y = X$ (i.e., Y always has the exact same value as X). In the questions below we use square loss function $\ell(a, x) = (a - x)^2$.

   (a) What is the Bayes risk?

   (b) What is the approximation error when using the hypothesis space of constant functions?

   (c) Suppose we use the hypothesis space $\mathcal{F}$ of affine functions.

      i. What is the approximation error?

      ii. Consider the function $\hat{f}(x) = x + 1$. Compute $R(\hat{f}) - R(f_{\mathcal{F}})$.

2. ($\star$) Let $\mathcal{X} = [-10, 10]$, $\mathcal{Y} = \mathcal{A} = \mathbb{R}$ and suppose the data distribution has marginal distribution $X \sim \text{Unif}(-10, 10)$ and $Y|X = x \sim \mathcal{N}(a + bx, 1)$. Throughout we assume the square loss function $\ell(a, x) = (a - x)^2$.

   (a) What is the Bayes risk?

   (b) What is the approximation error when using the hypothesis space of constant functions (in terms of a and b)?

(c) Suppose we use the hypothesis space of affine functions.

  i. What is the approximation error?

  ii. Suppose you have a fixed data set and compute the empirical risk minimizer $\hat{f}_n(x) = c + dx$. What is the estimation error (int terms of a, b, c, d) ?

3. Try to best characterize each of the following in terms of one or more of optimization error, approximation error, and estimation error.

   (a) Overfitting.

   (b) Underfitting.

   (c) Precise empirical risk minimization for your hypothesis space is computationally intractable.

   (d) Not enough data.

4. (a) We sometimes look at $R(\hat{f}_n)$ as random, and other times as deterministic. What causes this difference?

   (b) True or False: Increasing the size of our hypothesis space can shift risk from approximation error to estimation error but always leaves the quantity $R(\hat{f}_n) - R(f^*)$ constant.

   (c) True or False: Assume we treat our data set as a random sample and not a fixed quantity. Then the estimation error and the approximation error are random and not deterministic.

   (d) True or False: The empirical risk of the ERM, $\hat{R}(\hat{f}_n)$, is an unbiased estimator of the risk of the ERM $R(\hat{f}_n)$.

   (e) In each of the following situations, there is an implicit sample space in which the given expectation is computed. Give that space.

      i. When we say the empirical risk $\hat{R}(f)$ is an unbiased estimator of the risk $R(f)$ (where f is independent of the training data used to compute the empirical risk).

      ii. When we compute the expected empirical risk $\mathbb{E}[R(\hat{f}_n)]$ (i.e., the outer expectation).

      iii. When we say the minibatch gradient is an unbiased estimator of the full training set gradient.

5. For each, use $\leq$, $\geq$, or $=$ to determine the relationship between the two quantities, or if the relationship cannot be determined. Throughout assume $\mathcal{F}_1, \mathcal{F}_2$ are hypothesis spaces with $\mathcal{F}_1 \subseteq \mathcal{F}_2$, and assume we are working with a fixed loss function ℓ.

   (a) The estimation errors of two decision functions f_1, f_2 that minimize the empirical risk over the same hypothesis space, where f_2 uses 5 extra data points.

   (b) The approximation errors of the two decision functions f_1, f_2 that minimize risk with respect to $\mathcal{F}_1, \mathcal{F}_2$, respectively (i.e., $f_1 = f_{\mathcal{F}_1}$ and $f_2 = f_{\mathcal{F}_2}$).

   (c) The empirical risks of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively. Both use the same fixed training data.

   (d) The estimation errors (for $\mathcal{F}_1, \mathcal{F}_2$, respectively) of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively.

   (e) The risk of two decision functions f_1, f_2 that minimize the empirical risk over $\mathcal{F}_1, \mathcal{F}_2$, respectively.

6. In the excess risk decomposition lecture, we introduced the decision tree classifier spaces $\mathcal{F}$ (space of all decision trees) and $\mathcal{F}_d$ (the space of decision trees of depth d) and went through some examples. The following questions are based on those slides. Recall that $P_{\mathcal{X}} = \text{Unif}([0,1]^2)$, $\mathcal{Y} = \{\text{blue}, \text{orange}\}$, orange occurs with .9 probability below the line $y = x$ and blue occurs with .9 probability above the line $y = x$.

   (a) Prove that the Bayes error rate is 0.1.

(b) Is the Bayes decision function in $\mathcal{F}$?

(c) For the hypothesis space $\mathcal{F}_3$ the slide states that $R(\tilde{f}) = 0.176 \pm .004$ for $n = 1024$. Assuming you had access to the training code that produces $\tilde{f}$ from a set of data points, and random draws from the data generating distribution, give an algorithm (pseudocode) to compute (or estimate) the values 0.176 and .004.

2.2 Topic 2: Stochastic Gradient Descent

2.2.1 Learning Objectives

1. Be able to write the empirical risk for a particular loss function over a particular parameterized hypothesis space, such as for square loss over a hypothesis space of linear functions.

2. Compare and constrast gradient descent, minibatch gradient descent, and stochastic gradient descent.

2.2.2 Concept Check Questions

1. 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 $w \in \mathbb{R}^d$.

2. 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 $\epsilon > 0$?

3. ($\star$) 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.