Foundations of Machine Learning

Brett Bernstein

August 22, 2018

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 $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ is given by
$$\ell(a, y) = \mathbf{1}(a \neq y),$$
the $0-1$ loss.

Solution. The Bayes decision function f^* satisfies

$$f^* = \arg\min_f R(f) = \arg\min_f \mathbb{E}[\mathbf{1}(f(X) \neq Y)] = \arg\min_f P(f(X) \neq Y),$$

where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$. Let

$$f_1(x) = \arg\max_y P(Y = y \mid X = x),$$

the maximum a posteriori estimate of Y. If there is a tie, we choose any of the maximizers. If f_2 is another decision function we have

$$\begin{array}{rcll} P(f_1(X) \neq Y) & = & \sum_x P(f_1(x) \neq Y | X = x) P(X = x) & \\ & = & \sum_x (1 - P(f_1(x) = Y | X = x)) P(X = x) & \\ & \leq & \sum_x (1 - P(f_2(x) = Y | X = x)) P(X = x) & \text{(Defn of } f_1\text{)} \\ & = & \sum_x P(f_2(x) \neq Y | X = x) P(X = x) & \\ & = & P(f_2(X) \neq Y). & \end{array}$$

Thus $f^* = f_1$.

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?

   Solution. From Homework 1 we know that the Bayes decision function is given by $f^*(x) = \mathbb{E}[Y|X = x]$. Thus the Bayes risk is given by

   $$\mathbb{E}[(f^*(X) - Y)^2] = \mathbb{E}[(\mathbb{E}[Y|X] - Y)^2] = \mathbb{E}[\mathbb{E}[(\mathbb{E}[Y|X] - Y)^2|X]] = \mathbb{E}[\mathrm{Var}(Y|X)],$$

   where we applied the law of iterated expectations. The law of total variance states that

   $$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathbb{E}(Y|X)].$$

   This proves the Bayes risk satisfies

   $$\mathbb{E}[\mathrm{Var}(Y|X)] = \mathrm{Var}(Y) - \mathrm{Var}[\mathbb{E}(Y|X)] \leq \mathrm{Var}(Y).$$

   Recall from Homework 1 that $\mathrm{Var}(Y)$ is the Bayes risk when we estimate Y without any input X. This shows that using X in our estimation reduces the Bayes risk, and that the improvement is measured by $\mathrm{Var}[\mathbb{E}(Y|X)]$. As a sanity check, note that if X, Y are independent then $\mathbb{E}(Y|X) = \mathbb{E}(Y)$ so $\mathrm{Var}[\mathbb{E}(Y|X)] = 0$. If $X = Y$ then $\mathbb{E}(Y|X) = Y$ and $\mathrm{Var}[\mathbb{E}(Y|X)] = \mathrm{Var}(Y)$.

   The prominent role of variance in our analysis above is due to the fact that we are using the square loss.

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 $X \sim \mathrm{Unif}\{1, \ldots, 10\}$ and conditional distribution $Y|X = x \sim \mathrm{Unif}\{1, \ldots, x\}$. For each loss function below give a Bayes decision function.

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

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

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

Solution.

(a) From Homework 1 we know that $f^*(x) = \mathbb{E}[Y|X = x] = (x + 1)/2$.

(b) From Homework 1, we know that $f^*(x)$ is the conditional median of Y given $X = x$. If x is odd, then $f^*(x) = (x + 1)/2$. If x is even, then we can choose any value in the interval

$$\left[\left\lfloor \frac{x+1}{2} \right\rfloor, \left\lceil \frac{x+1}{2} \right\rceil\right].$$

(c) From question 1 above, we know that $f^*(x) = \arg\max_y P(Y = y|X = x)$. Thus we can choose any integer between 1 and x, inclusive, for $f^*(x)$.

4. Show that the empirical risk is an unbiased and consistent estimator of the Bayes risk. You may assume the Bayes risk is finite.

Solution. We assume a given loss function ℓ and an i.i.d. sample $(x_1, y_i), \ldots, (x_n, y_n)$. To show it is unbiased, note that

$$\begin{aligned}
\mathbb{E}[\hat{R}_n(f)] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\ell(f(x_i), y_i)] \qquad \text{(Linearity of } \mathbb{E}\text{)} \\
&= \mathbb{E}[\ell(f(x_1), y_1)] \qquad \text{(i.i.d.)} \\
&= R(f).
\end{aligned}$$

For consistency, we must show that as $n \to \infty$ we have $\hat{R}_n(f) \to R(f)$ with probability 1. Letting $z_i = \ell(f(x_i), y_i)$, we see that the z_i are i.i.d. with finite mean. Thus consistency follows by applying the strong law of large numbers.

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?

Solution.

(a) We can let $\hat{f}(x) = 5$ or $\hat{f}(x) = 3$ and obtain the minimal empirical risk of $3/5$. Thus the empirical risk minimizer is not unique.

(b) One solution is to let $\hat{f}(x) = 5$ for $x \in [0, .1]$ and $\hat{f}(x) = 3$ for $x \in (.1, 1]$ giving an empirical risk of $2/5$. There are uncountably many empirical risk minimizers, so again we do not have uniqueness.

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.

Solution.

(a) For any decision function f the risk is given by

$$\mathbb{E}[\mathbf{1}(f(X) \neq Y)] = P(f(X) \neq Y) = 1 - P(f(X) = Y) = 1.$$

To see this note that

$$P(f(X) = Y) = \frac{1}{20\sqrt{2\pi}} \int_{-10}^{10} \int_{-\infty}^{\infty} \mathbf{1}(f(x) = y) e^{-(y-a-bx)^2/2} \, dy \, dx = \frac{1}{20\sqrt{2\pi}} \int_{-10}^{10} 0 \, dx = 0.$$

Thus every decision function is a Bayes decision function, and the Bayes risk is 1.

(b) By problem 2 above we know the Bayes risk is given by

$$\mathbb{E}[\text{Var}(Y|X)] = \mathbb{E}[1] = 1,$$

since $\text{Var}(Y|X = x) = 1$.

(c) We choose $\hat{f}$ such that

$$\hat{f}(0) = 1.5, \hat{f}(1) = 3, \hat{f}(2.5) = 3.1, \hat{f}(-4) = 2.1,$$

and $\hat{f}(x) = 0$ otherwise. Then we achieve the minimum empirical risk of $1/10$.

(d) Letting

$$A = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2.5 \\ 1 & -4 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 3.1 \\ -2.1 \end{pmatrix}$$

we obtain (using a computer)

$$\hat{w} = \begin{pmatrix} \hat{d} \\ \hat{c} \end{pmatrix} = (A^T A)^{-1} A^T y = \begin{pmatrix} 1.4856 \\ 0.8556 \end{pmatrix}.$$

This gives

$$\hat{R}_5(\hat{f}) = \frac{1}{5}\|A\hat{w} - y\|_2^2 = 0.2473.$$

[Aside: In general, to solve systems like the one above on a computer you shouldn't actually invert the matrix $A^T A$, but use something like w=A\y in Matlab which performs a QR factorization of A.]

(e) Letting

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2.5 & 6.25 \\ 1 & -4 & 16 \end{pmatrix}, \quad y = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 3.1 \\ -2.1 \end{pmatrix}$$

we obtain (using a computer)

$$\hat{w} = \begin{pmatrix} \hat{e} \\ \hat{d} \\ \hat{c} \end{pmatrix} = (A^T A)^{-1} A^T y = \begin{pmatrix} 1.7175 \\ 0.7545 \\ -0.0521 \end{pmatrix}.$$

This gives

$$\hat{R}_5(\hat{f}) = \frac{1}{5}\|A\hat{w} - y\|_2^2 = 0.1928.$$

Topic 2: Stochastic Gradient Descent

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.

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

   Solution. Let $(x_{m_1}, y_{m_1}), \ldots, (x_{m_n}, y_{m_n})$ be our mini-batch selected uniformly without replacement from the full training set $(x_1, y_1), \ldots, (x_n, y_n)$.

$$\begin{aligned}
\mathbb{E}\left[\nabla_w \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_{m_i}, y_{m_i}))\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\nabla_w \ell(f_w(x_{m_i}), y_{m_i})\right] && \text{(Linearity of } \nabla, \mathbb{E}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\nabla_w \ell(f_w(x_{m_1}), y_{m_1})\right] && \text{(Marginals are the same)} \\
&= \mathbb{E}\left[\nabla_w \ell(f_w(x_{m_1}), y_{m_1})\right] \\
&= \sum_{i=1}^{N} \frac{1}{N} \nabla_w \ell(f_w(x_i), y_i) \\
&= \nabla_w \frac{1}{N}\sum_{i=1}^{N} \ell(f_w(x_i), y_i) && \text{(Linearity of } \nabla).
\end{aligned}$$

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 $\epsilon > 0$?

   Solution. Let $x_1, \ldots, x_n$ denote our randomly sampled ages, and let $\hat{x}$ denote the sample mean $\frac{1}{n}\sum_{i=1}^{n} x_i$. Then

$$\mathrm{Var}(\hat{x}) = \frac{\sigma^2}{n}.$$

   Thus we require $n \geq \sigma^2/\epsilon$. Note that this doesn't depend on N, the full population size.

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.

Solution. Throughout we assume our gradient lies in $\mathbb{R}^d$. Consider the empirical distribution on the full training set (i.e., each sample is chosen with probability $1/N$ where N is the full training set size). Assume this distribution has mean vector $\mu \in \mathbb{R}^d$ (the full-batch gradient) and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. By the central limit theorem the mini-batch gradient will be approximately normally distributed with mean μ and covariance $\frac{1}{n}\Sigma$, where n is the mini-batch size. As N grows the entries of Σ need not grow, and thus n need not grow. In fact, as N grows, the empirical mean and covariance matrix will converge to their true values. More precisely, the mean of the empirical distribution will converge to $\mathbb{E}\nabla\ell(f(X), Y)$ and the covariance will converge to

$$\mathbb{E}[(\nabla\ell(f(X), Y))(\nabla\ell(f(X), Y))^T] - \mathbb{E}[\nabla\ell(f(X), Y)]\mathbb{E}[\nabla\ell(f(X), Y)]^T$$

where $(X, Y) \sim P_{\mathcal{X}\times\mathcal{Y}}$.

The important takeaway here is that the size of the mini-batch is dependent on the speed of computation, and on the characteristics of the distribution of the gradients (such as the moments), and thus may vary independently of the size of the full training set.