Machine Learning – Brett Bernstein

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 $\ell : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ 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 $X \sim \mathrm{Unif}\{1, \ldots, 10\}$ 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) $\ell(a, y) = \mathbf{1}(a \neq y)$.

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.

Stochastic Gradient Descent

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.

Week 1 Lab: Concept Check Exercises

Starred problems are optional.

Multivariable Calculus Exercises

1. If $f'(x;u) < 0$ show that $f(x+hu) < f(x)$ for sufficiently small $h > 0$.

2. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, and assume that $\nabla f(x) \neq 0$. Prove

$$\arg\max_{\|u\|_2=1} f'(x;u) = \frac{\nabla f(x)}{\|\nabla f(x)\|_2} \quad \text{and} \quad \arg\min_{\|u\|_2=1} f'(x;u) = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2}.$$

3. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x,y) = x^2 + 4xy + 3y^2$. Compute the gradient $\nabla f(x,y)$.

4. Compute the gradient of $f : \mathbb{R}^n \to \mathbb{R}$ where $f(x) = x^T A x$ and $A \in \mathbb{R}^{n \times n}$ is any matrix.

5. Compute the gradient of the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by

$$f(x) = b + c^T x + x^T A x,$$

   where $b \in \mathbb{R}$, $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$.

6. Fix $s \in \mathbb{R}^n$ and consider $f(x) = (x-s)^T A (x-s)$ where $A \in \mathbb{R}^{n \times n}$. Compute the gradient of f.

7. Consider the ridge regression objective function

$$f(w) = \|Aw - y\|_2^2 + \lambda \|w\|_2^2,$$

   where $w \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$, and $\lambda \in \mathbb{R}_{\geq 0}$.

   (a) Compute the gradient of f.
   (b) Express f in the form $f(w) = \|Bw - z\|_2^2$ for some choice of B, z.
   (c) Using either of the parts above, compute

$$\arg\min_{w \in \mathbb{R}^n} f(w).$$

8. Compute the gradient of

$$f(\theta) = \lambda \|\theta\|_2^2 + \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i)),$$

   where $y_i \in \mathbb{R}$ and $\theta \in \mathbb{R}^m$ and $x_i \in \mathbb{R}^m$ for $i = 1, \ldots, n$.

Linear Algebra Exercises

1. When performing linear regression we obtain the *normal equations* $A^T Ax = A^T y$ where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$, and $y \in \mathbb{R}^m$.
   (a) If $\mathbf{rank}(A) = n$ then solve the normal equations for x.
   (b) ($\star$) What if $\mathbf{rank}(A) \neq n$?
2. Prove that $A^T A + \lambda \mathbf{I}_{n \times n}$ is invertible if $\lambda > 0$ and $A \in \mathbb{R}^{n \times n}$.
3. ($\star$) Describe the following set geometrically:

$$\left\{ v \in \mathbb{R}^2 \;\middle|\; v^T \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix} v = 4 \right\}.$$

Week 2 Pre-Lecture: Concept Check Exercises

Optimization Prerequisites for Lasso

L1 and L2 Regularization

1. Given $a \in \mathbb{R}$ we define a^+, a^- as follows:

$$a^+ = \begin{cases} a & \text{if } a \geq 0, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{and} \quad a^- = \begin{cases} -a & \text{if } a < 0, \\ 0 & \text{otherwise.} \end{cases}$$

   We call a^+ the *positive part* of a and a^- the *negative part* of a. Note that $a^+, a^- \geq 0$.
   (a) Give an expression for a in terms of a^+, a^-.
   (b) Give an expression for $|a|$ in terms of a^+, a^-.
   For $x \in \mathbb{R}^d$ define $x^+ = (x_1^+, \dots, x_d^+)$ and $x^- = (x_1^-, \dots, x_d^-)$.
   (c) Give an expression for x in terms of x^+, x^-.
   (d) Give an expression for $\|x\|_1$ without using any summations or absolute values. [Hint: Use x^+, x^- and the vector $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{R}^d$.]
2. Let $f : \mathbb{R} \to \mathbb{R}$ and $S \subseteq \mathbb{R}$. Consider the two optimization problems

$$\begin{array}{ll} \text{minimize}_{x \in \mathbb{R}} & |x| \\ \text{subject to} & f(x) \in S \end{array} \qquad \text{and} \qquad \begin{array}{ll} \text{minimize}_{a,b \in \mathbb{R}} & a + b \\ \text{subject to} & f(a - b) \in S \\ & a, b \geq 0. \end{array}$$

   Solve the following questions.
   (a) If x in the first problem satistfies $f(x) \in S$ show how to quickly compute (a, b) for the second problem with $a + b = |x|$ and $f(a - b) \in S$.

(b) If a, b in the second problem satisfy $f(a-b) \in S$, show how to quickly compute an x for the first problem with $|x| \leq a+b$ and $f(x) \in S$.

(c) Assume x is a minimizer for the first problem with minimum value p_1^* and (a, b) is a minimizer for the second problem with minimum p_2^*. Using the previous two parts, conclude that $p_1^* = p_2^*$.

3. Let $f : \mathbb{R}^d \to \mathbb{R}$, $S \subseteq \mathbb{R}$ and consider the following optimization problem:

$$\begin{array}{ll} \text{minimize}_{x \in \mathbb{R}^d} & \|x\|_1 \\ \text{subject to} & f(x) \in S, \end{array}$$

where $\|x\|_1 = \sum_{i=1}^d |x_i|$. Give a new optimization problem with a linear objective function and the same minimum value. Show how to convert a solution to your new problem into a solution to the given problem. [Hint: Use the previous two problems.]