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.

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.

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

Lab 1: Gradients and Directional Derivatives

Multivariate Differentiation

Learning Objectives

1. Define the directional derivative, and use it to find a linear approximation to $f(\mathbf{x}+h\mathbf{u})$.
2. Define partial derivative and the gradient. Show how to compute an arbitrary directional derivative using the gradient.
3. For a differentiable function, give a linear approximation near a point $\mathbf{x}$ using the gradient.
4. Show that the gradient gives the direction of steepest ascent, and the negative gradient gives the direction of steepest descent.

Concept Check Questions

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

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

Computing Gradients

Learning Objectives

1. Find the gradient of a function by computing each partial derivative separately.
2. Use the chain rule to perform gradient computations.
3. Compute the gradient of a differentiable function by determining the form of a general directional derivative.

Concept Check Questions

1. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x,y) = x^2 + 4xy + 3y^2$. Compute the gradient $\nabla f(x,y)$.
2. Compute the gradient of $f : \mathbb{R}^n \to \mathbb{R}$ where $f(x) = x^T Ax$ and $A \in \mathbb{R}^{n\times n}$ is any matrix.
3. Compute the gradient of the quadratic function $f : \mathbb{R}^n \to \mathbb{R}$ given by

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

   where $b \in \mathbb{R}$, $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$.
4. 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.
5. 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. What do you notice about B?

   (c) Using either of the parts above, compute

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

6. 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 $i = 1, \ldots, n$.

Pre-Lecture 2: Optimization and linear algebra

Instructions: Prior to lecture 2, please review the following problems

Optimization Prerequisites for Lasso

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

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

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 $x^+ = (x_1^+, \ldots, x_d^+)$ and $x^- = (x_1^-, \ldots, x_d^-)$.

(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 $\mathbf{1} = (1, 1, \ldots, 1) \in \mathbb{R}^d$.]

2. Let $f : \mathbb{R} \to \mathbb{R}$ and $S \subseteq \mathbb{R}$. Consider the two optimization problems

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

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.]

Ellipsoids

1. ($\star$) Describe the following set geometrically:

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

($\star$) Linear Algebra Prerequisites for Linear Regressions

1. When performing linear regression we obtain the *normal equations* $A^T A x = A^T y$ 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 $A \in \mathbb{R}^{n \times n}$.

Lecture 2: Excess Risk Decomposition and Regularization

Topic 1: Excess Risk Decomposition

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?

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.

Topic 2: L_1 and L_2 Regularization

Learning Objectives

1. Explain the concept of a sequence of nested hypothesis spaces, and explain how a complexity measure (of a function) can be used to create such a sequence.

2. Given a base hypothesis space of decision functions (e.g. affine functions), a performance measure for a decision function (e.g. empirical risk on a training set), and a function complexity measure (e.g. Lipschitz continuity constant of decision function), give the corresponding optimization problem in Tikhonov and Ivanov forms.

3. For some situations (i.e. combinations of base hypothesis space, performance measure, and complexity measure), we claimed that Tikhonov and Ivanov forms are equivalent. Be able to explain what this means and write it down mathematically.

4. In particular, the Tikhonov and Ivanov formulations are equivalent for lasso and ridge regression. Be comfortable switching between the formulations to assist with interpretations (e.g. the classic L1 regularization picture with the norm ball is based on the Ivanov formulation).

Concept Check Questions

1. Consider the following two minimization problems:

$$\arg\min_w \Omega(w) + \frac{\lambda}{n}\sum_{i=1}^{n} L(f_w(x_i), y_i)$$

and

$$\arg\min_w C\Omega(w) + \frac{1}{n}\sum_{i=1}^{n} L(f_w(x_i), y_i),$$

where $\Omega(w)$ is the penalty function (for regularization) and L is the loss function. Give sufficient conditions under which these two give the same minimizer.

2. ($\star$) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Prove that $\|\nabla f(x)\|_2 \leq L$ if and only if f is Lipschitz with constant L.

3. ($\star$) Let $\hat{w}$ denote the minimizer for

$$\begin{array}{ll} \text{minimize}_w & \|Xw - y\|_2^2 \\ \text{subject to} & \|w\|_1 \leq r. \end{array}$$

Prove that $f(x) = \hat{w}^T x$ is Lipschitz with constant r.

4. Two of the plots in the lecture slides use the fact that $\|\hat{w}\|/\|\tilde{w}\|$ is always between 0 and 1. Here $\hat{w}$ is the parameter vector of the linear model resulting from the regularized least squares problem. Analgously, $\tilde{w}$ is the parameter vector from the unregularized problem. Why is this true that the quotient lies in $[0, 1]$?

5. Explain why feature normalization is important if you are using L_1 or L_2 regularization.

Week 4 Lab: Concept Check Exercises

Subgradients

1. ($\star$) If $f : \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable at x, the $\partial f(x) = \{\nabla f(x)\}$.

2. Fix $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathbb{R}^n$. Then the subdifferential $\partial f(x)$ is a convex set.

3. (a) True or False: A subgradient of $f : \mathbb{R}^n \to \mathbb{R}$ at x is normal to a hyperplane that globally understimates the graph of f.
   (b) True or False: If $g \in \partial f(x)$ then $-g$ is a descent direction of f.
   (c) True or False: For $f : \mathbb{R} \to \mathbb{R}$, if $1, -1 \in \partial f(x)$ then x is a global minimizer of f.
   (d) True or False: Let $f : \mathbb{R}^n \to \mathbb{R}$ and let $g \in \partial f(x)$. Then $\alpha g \in \partial f(x)$ for all $\alpha \in [0, 1]$.
   (e) True or False: If the sublevel sets of a function are convex, then the function is convex.

4. Let $f : \mathbb{R}^2 \to \mathbb{R}$ be defined by $f(x_1, x_2) = |x_1| + 2|x_2|$. Compute $\partial f(x_1, x_2)$ for each $x_1, x_2 \in \mathbb{R}^2$.

Lecture 4: Concept Checks

Convexity

Optional Learning Objectives

Convex optimization and Lagrangian duality will not covered on the midterm exam, so in some sense these objectives are optional.

- Define a convex set, a convex function, and a strictly convex function. (Don't forget that the domain of a convex function must be a convex set!)

- For an optimization problem, define the terms feasible set, feasible point, active constraint, optimal value, and optimal point.

- Give the form for a general inequality-constrained optimization problem (there are many ways to do this, but our convention is to have inequality constraints of the form $f_i(x) \leq 0$).
- Define the Lagrangian for this optimization problem, and explain how the Lagrangian encodes all the information in the original optimization problem.
- Write the primal and dual optimization problem in terms of the Lagrangian.

Convexity Concept Check Problems

1. If $A, B \subseteq \mathbb{R}^n$ are convex, then $A \cap B$ is convex.
2. Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be convex. Show that $af + bg$ is convex if $a, b \geq 0$.
3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Prove that if $\nabla f(x) = 0$ then x is a global minimizer.
4. Prove that if $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex and x is a global minimizer, then it is the unique global minimizer.
5. Prove that any affine function $f : \mathbb{R}^n \to \mathbb{R}$ is both convex and concave.
6. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and let $g : \mathbb{R}^m \to \mathbb{R}^n$ be affine. Then $f \circ g$ is convex.
7. ($\star\star$)
   (a) Let $f : \mathbb{R} \to \mathbb{R}$ be convex. Show that f has one-sided left and right derivatives at every point.
   (b) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Show that f has one-sided directional derivatives at every point.
   (c) Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. Show that if x is not a minimizer of f then f has a descent direction at x (i.e., a direction whose corresponding one-sided directional derivative is negative).

Convex Optimization Problems

1. Suppose there are mn people forming m rows with n columns. Let a denote the height of the tallest person taken from the shortest people in each column. Let b denote the height of the shortest person taken from the tallest people in each row. What is the relationship between a and b?
2. Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be given data. You want to find the center and radius of the smallest sphere that encloses all of the points. Express this problem as a convex optimization problem.

3. Suppose $x_1, \ldots, x_n \in \mathbb{R}^d$ and $y_1, \ldots, y_n \in \{-1, 1\}$. Here we look at y_i as the label of x_i. We say the data points are linearly separable if there is a vector $v \in \mathbb{R}^d$ and $a \in \mathbb{R}$ such that $v^T x_i > a$ when $y_i = 1$ and $v^T x_i < a$ for $y_i = -1$. Give a method for determining if the given data points are linearly separable.

4. Consider the Ivanov form of ridge regression:

$$\begin{array}{ll} \text{minimize} & \|Ax - y\|_2^2 \\ \text{subject to} & \|x\|_2^2 \leq r^2, \end{array}$$

where $r > 0$, $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$ are fixed.

   (a) What is the Lagrangian?

   (b) What do you get when you take the supremum of the Lagrangian over the feasible values for the dual variables?

Week 5 Lab: Concept Check Exercises

Kernels

Kernel Learning Objectives

- Explain how explicit feature maps can be used to extend the expressivity of linear models.
- Explain potential issues explicitly computing large feature spaces.
- State and explain the definition of a 'kernelized' method.
- Explain why the SVM dual is kernelized, while the primal is not (ignoring the representer theorem).
- Give the relationship between a feature map and kernel function.
- Explain the computational benefits of kernelization based on costs of optimizing over $\mathbb{R}^n$ vs $\mathbb{R}^d$.
- Be able to apply the kernel trick using the kernel matrix K.
- Be able to apply the elements of our proof of the representer theorem (ex: projections decrease norms) to prove related theorems.
- Compare using the representer theorem and duality to kernelized SVM.
- Describe common kernels (RBF/polynomial) and their properties (i.e. equivalent feature maps, computational benefits relative to explicit computation (if possible),...).
- Describe some general recipes for deriving "new" kernel function.

Kernel Concept Check Questions

1. Fix $n > 0$. For $x, y \in \{1, 2, \ldots, n\}$ define $k(x, y) = \min(x, y)$. Give an explicit feature map $\varphi : \{1, 2, \ldots, n\}$ to $\mathbb{R}^D$ (for some D) such that $k(x, y) = \varphi(x)^T \varphi(y)$.
2. Show that $k(x, y) = (x^T y)^4$ is a positive semidefinite kernel on $\mathbb{R}^d \times \mathbb{R}^d$.
3. Let $A \in \mathbb{R}^{d \times d}$ be a positive semidefinite matrix. Prove that $k(x, y) = x^T A y$ is a positive semidefinite kernel.
4. Consider the objective function

$$J(w) = \|Xw - y\|_1 + \lambda \|w\|_2^2.$$

   Assume we have a positive semidefinite kernel k.

   (a) What is the kernelized version of this objective?

   (b) Given a new test point x, find the predicted value.
5. Show that the standard 2-norm on $\mathbb{R}^n$ satisfies the parallelogram law.
6. Suppose you are given an training set of distinct points $x_1, x_2, \ldots, x_n \in \mathbb{R}^n$ and labels $y_1, \ldots, y_n \in \{-1, +1\}$. Show that by properly selecting σ you can achieve perfect $0-1$ loss on the training data using a linear decision function and the RBF kernel.
7. Suppose you are performing standard ridge regression, which you have kernelized using the RBF kernel. Prove that any decision function $f_\alpha(x)$ learned on a training set must satisfy $f_\alpha(x) \to 0$ as $\|x\|_2 \to \infty$.
8. Consider the standard (unregularized) linear regression problem where we minimize $L(w) = \|Xw - y\|_2^2$ for some $X \in \mathbb{R}^{n \times m}$ and $y \in \mathbb{R}^n$. Assume $m > n$.

   (a) Let w^* be one minimizer of the loss function L above. Give an infinite set of minimizers of the loss function.

   (b) What property defines the minimizer given by the representer theorem (in terms of X)?

Multiclass: Concept Check

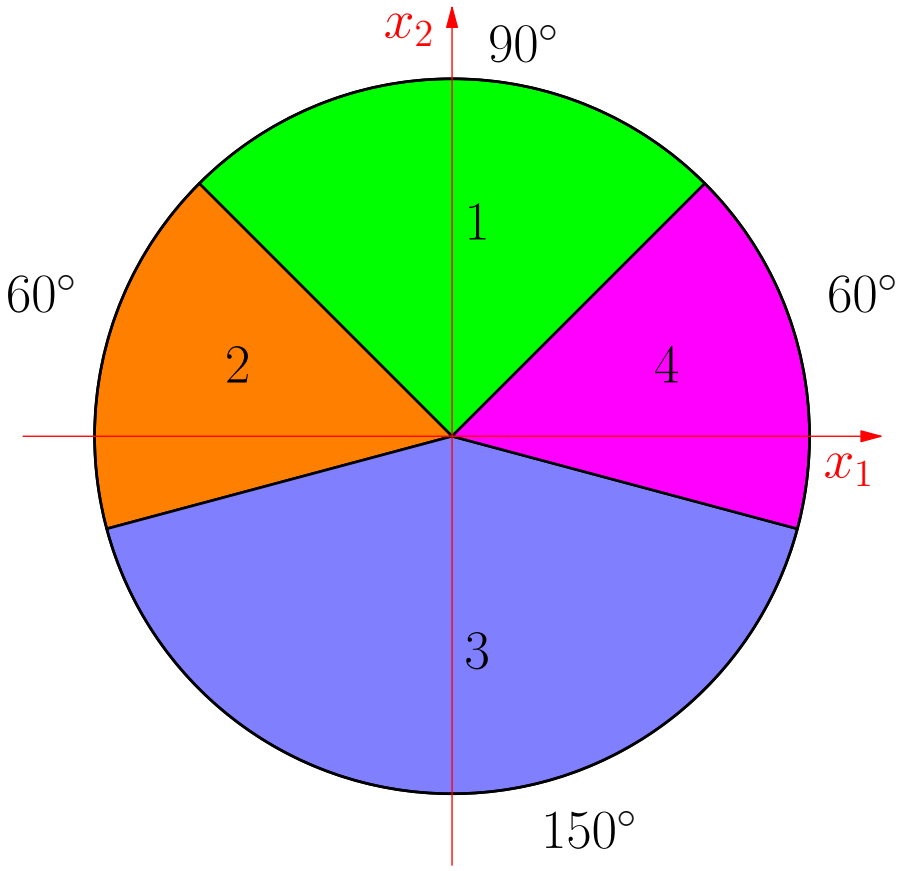
Multiclass Learning Objectives

- Be able to give pseudocode to fit and apply a one-vs-all/one-vs-rest prediction function.
- Be able to describe an example where one-vs-all fails.
- Be able to explain our reframing of multiclass learning in terms of a compatability score function.

- Be able to define the class-specific margin of a data instance using the compatability score function.
- Be able to map a set of linear score functions onto a single linear class-sensitive score function using a class-sensitive feature map. Give some intuition for the value of this feature map (based on features related to the target classes).
- Be able to state the multiclass SVM objective with 1 as the target margin, and be able to generalize using a class-specific target-margin and explain this generalization using the intuition of this target-margin as a lookup table.

Multiclass Concept Check Questions

1. Let $\mathcal{X} = \mathbb{R}^2$ and $\mathcal{Y} = \{1, 2, 3, 4\}$, with X uniformly distributed on $\{x \mid \|x\|_2 \leq 1\}$. Given X, the value of Y is determined according to the following image, where green is 1, orange is 2, blue is 3, and magenta is 4.



For the problems below we are using the 0-1 loss.

(a) Consider the multiclass linear hypothesis space

$$\mathcal{F} = \{f \mid f(x) = \underset{i \in \{1,2,3,4\}}{\arg\max}\, w_i^T x\},$$

where each f is determined by $w_1, w_2, w_3, w_4 \in \mathbb{R}^2$. Give $f_{\mathcal{F}}$, a decision function minimizing the risk over $\mathcal{F}$, by specifying the corresponding w_1, w_2, w_3, w_4. Then give $R(f_{\mathcal{F}})$.

(b) Now consider the restricted hypothesis space

$$\mathcal{F}_1 = \{f \mid f(x) = \underset{i \in \{1,2,3,4\}}{\arg\max}\, w_i^T x, \|w_1\| = \|w_2\| = \|w_3\| = \|w_4\| = 1\}.$$

Consider the decision function $f \in \mathcal{F}_1$ with w_1, w_2, w_3, w_3 set to the angle bisectors of the corresponding regions. Give $R(f)$.

(c) Next consider the class-sensitive version of $\mathcal{F}$:

$$\mathcal{F}_2 = \{f \mid f(x) = \underset{i \in \{1,2,3,4\}}{\arg\max}\, w^T \Psi(x, i)\},$$

where $w \in \mathbb{R}^D$ and $\Psi : \mathbb{R}^2 \times \{1, 2, 3, 4\} \to \mathbb{R}^D$. Give w, Ψ corresponding to $f_{\mathcal{F}_2}$, the decision function minimizing the risk over $\mathcal{F}_2$.

2. Recall that the standard (featurized) SVM objective is given by

$$J_1(w) = \frac{1}{2}\|w\|_2^2 + \frac{C}{n}\sum_{i=1}^{n}[1 - y_i w^T \varphi(x_i)]_+.$$

The 2-class multiclass SVM objective is given by

$$J_2(w) = \frac{1}{2}\|w\|_2^2 + \frac{C}{n}\sum_{i=1}^{n}\max_{y \neq y_i}[1 - m_{i,y}(w)]_+,$$

where $m_{i,y}(w) = w^T\Psi(x_i, y_i) - w^T\Psi(x_i, y)$. Give a Ψ (in terms of φ) so that multiclass with 2 classes $\{-1, +1\}$ is equivalent to our standard SVM objective.

3. Suppose you trained a decision function f from the hypothesis space $\mathcal{F}$ given by

$$\mathcal{F} = \{f \mid f(x) = \underset{i \in \{1,\ldots,k\}}{\arg\max}\, w^T \psi(x, i)\}.$$

Give pseudocode showing how you would use f to forecast the class of a new data point x.

4. Consider a multiclass SVM with objective

$$J(w) = \frac{1}{2}\|w\|_2^2 + \frac{C}{n}\sum_{i=1}^{n}\max_{y \neq y_i}[1 - m_{i,y}(w)]_+,$$

where $m_{i,y}(w) = w^T\Psi(x_i, y_i) - w^T\Psi(x_i, y)$. Assume $\mathcal{Y} = \{1, \ldots, k\}$, $\mathcal{X} = \mathbb{R}^d$, $w \in \mathbb{R}^D$ and $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^D$. Give a kernelized version of the objective.

Trees, Bootstrap, Bagging, and RFs

Trees

Trees Learning Objectives

- Be able to describe the structure of a binary tree (ex: put bounds on number of leaves given height; describe the geometry of the resulting prediction function; etc.).
- Give pseudocode for finding the optimal split for (a) a continuous feature, and (b) a categorical feature for a binary classification problem.
- Describe some reasonable strategies for controlling the complexity of a tree.
- In particular, describe the regularization approach used in CART (pruning and use of number of leaves as complexity measure), recognize the cost complexity criterion as our standard regularized ERM.
- Recall the entropy, Gini, and misclassification error splitting criteria. Give some intuition around preference for Gini/entropy (i.e. purity measures) over misclassification.

Trees Concept Check Questions

1. (a) How many regions (leaves) will a tree with k node splits have?
   (b) What is the maximum number of regions a tree of height k can have? Recall that the height of a tree is the number of edges in the longest path from the root to any leaf.
   (c) Give an upper bound on the depth needed to exactly classify n distinct points in $\mathbb{R}^d$. [Hint: In the worst case each leaf will have a single training point.]
2. This question involves fitting a regression tree using the square loss. Assume the n data points for the current node are sorted by the first feature. Give pseudocode with $O(n)$ runtime for optimally splitting the current node with respect to the first feature.
3. Suppose we are looking at a fixed node of a classification tree, and the class labels are, sorted by the first feature values,

   $$4, 1, 0, 0, 1, 0, 2, 3, 3.$$

   We are currently testing splitting the node into a left node containing $4, 1, 0, 0, 1, 0$ and a right node containing $2, 3, 3$. For each of the following impurity measures, give the value for the left and right parts, along with the total score for the split.
   (a) Misclassification error.
   (b) Gini index.
   (c) Entropy.

Bootstrap and Bagging

Bootstrap and Bagging Learning Objectives

- Recall from basic statics concepts related to an estimator (e.g. bias) and its variance.
- Describe (outside the context of bagging/RFs) how the bootstrap is a useful method for estimating the variance of an estimator, and have some intuition on how it can be applied across many problems.
- Again recalling basic statistics, understand why bagging (averaging predictions) reduces variance.
- Recalling that the bootstrap ignores an expected 37% of data in each bootstrap sample, explain how we can use out-of-bag observations to approximate test performance.
- Describe how RF reduces correlation between trees using column sampling while training on bootstrap samples.

Bootstrap and Bagging Concept Check Questions

1. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a distribution with mean μ and variance σ^2. How large must n be so that the sample mean has standard error smaller than .01?
2. Let $X_1, \ldots, X_{2n+1}$ be an i.i.d. sample from a distribution. To estimate the median of the distribution, you can compute the sample median of the data.
   (a) Give pseudocode that computes an estimate of the variance of the sample median.
   (b) Give pseudocode that computes an estimate of a 95% confidence interval for the median.

Conditional Probability Models: Concept Check

Conditional Probability Models

MLE Learning Objectives

- Define the likelihood of an estimate of a probability distribution for some data $\mathcal{D}$.
- Define a parameteric model, and some common parameteric families.
- Define the MLE for some parameter θ of a probability model.
- Be able to find the MLE using first order conditions on the log-likelihood.

Conditional Probability Models

- Describe the basic structure of a linear probabilistic model, in terms of (i) a parameter θ of the probablistic model, (ii) a linear score function, (iii) a transfer function (kin to a "response function" or "inverse link" function, though we've relaxed requirements on the parameter theta).
- Explain how we can use MLE to choose w, the weight vector in our linear function (in (ii) above).
- Give common transfer functions for (i) bernoulli, (ii) poisson, (iii) gaussian, and (iv) categorical distributions. Explain why these common transfer functions make sense (in terms of their codomains).
- Explain the equivalence of EMR and MLE for negative log-likelihood loss.

MLE/Conditional Probability Model Concept Check Question

1. In each of the following, assume $X_1, \ldots, X_n$ are an i.i.d. sample from the given distribution.
   (a) Compute the MLE for p assuming each $X_i \sim \text{Geom}(p)$ with PMF $f_X(k) = (1-p)^{k-1}p$ for $k \in \mathbb{Z}_{\geq 1}$.
   (b) Compute the MLE for λ assuming each $X_i \sim \text{Exp}(\lambda)$ with PDF $f_X(x) = \lambda e^{-\lambda x}$.
2. We want to fit a regression model where $Y|X = x \sim \text{Unif}([0, e^{w^T x}])$ for some $w \in \mathbb{R}^d$. Given i.i.d. data points $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$, give a convex optimization problem that finds the MLE for w.
3. Explain why softmax is related to computing the maximum of a list of values.
4. Suppose x has a Poisson distribution with unknown mean θ:

$$p(x|\theta) = \frac{\theta^x}{x!}\exp(-\theta), \qquad x = 0, 1, \cdots$$

   Let the prior for θ be a gamma distribution:

$$p(\theta|\alpha, \beta) = \frac{\beta^\alpha \theta^{\alpha-1}}{\Gamma(\alpha)}\exp(-\beta\theta), \qquad \theta > 0$$

   where Γ is the gamma function. Show that, given an observation x, the posterior $p(\theta|x, \alpha, \beta)$ is a gamma distribution with updated parameters $(\alpha', \beta') = (\alpha + x, \beta + 1)$. What does this tell you about the Poisson and gamma distributions?

Bayesian Methods and Regression: Concept Check

Bayesian Methods and Regression

Bayesian Methods and Regression Learning Objectives

- (Recap) Recall the basic Bayesian setup (likelihood and prior), and be able to write the posterior distribution using proportionality – (see slide 15 for Gaussian Example).
- Explain the difference between the posterior predictive distribution function and the MAP or posterior mean estimator.
- Be able to show the relationship between Gaussian regression and ridge regression.
- Explain what a predictive distribution is, and how it gives additional information (relative to the prediction functions we've learned in our ridge/lasso homework, for example).

Bayesian Methods and Regression Concept Check Questions

1. (From DeGroot and Schervish) Let θ denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of θ is unknown, and two statisticians A and B assign to θ the following different prior PDFs $\xi_A(\theta)$ and $\xi_B(\theta)$, respectively:

$$\begin{aligned} \xi_A(\theta) &= 2\theta \quad \text{for } 0 < \theta < 1, \\ \xi_B(\theta) &= 4\theta^3 \quad \text{for } 0 < \theta < 1. \end{aligned}$$

   In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

   (a) Find the posterior distribution that each statistician assigns to θ.

   (b) Find the Bayes estimate of θ (minimizer of posterior expected loss) for each statistician based on the squared error loss function.

   (c) Show that after the opinions of the 1000 registered voters in the random sample had been obtained, the Bayes estimates for the two statisticians could not possibly differ by more than 0.002, regardless of the number in the sample who were in favor of the proposition.

2. Two statistics students decide to compute 95% confidence intervals for the distribution parameter θ using an i.i.d. sample $X_1, \ldots, X_n$. Student B uses Bayesian methods to find a 95% credible set $[L_B, R_B]$ for θ. Student F uses frequentist methods to find a 95% confidence interval $[L_F, R_F]$ for θ. Both conclude that parameter θ is in their respective intervals with probability at least .95. Who is correct? Explain.

3. Suppose θ has prior distribution Beta(a, b) for some $a, b > 0$. Given θ, suppose we make independent coin flips with heads probability θ. Find values of a, b and the coin flips so that the posterior variance is larger than the prior variance. [Hint: Recall that a Beta(a, b) random variable has variance given by

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

Try $b = 1$.]

4. Fix $\sigma^2 > 0$. Let w, taking values in $\mathbb{R}^d$, have prior distribution $\mathcal{N}(\mu_0, \Sigma_0)$. Conditional on w and $x_1, \ldots, x_n \in \mathbb{R}^2$ suppose that $y_1, \ldots, y_n$ are i.i.d. with $y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$. Let $\mathcal{N}(\mu_1, \Sigma_1)$ denote the posterior distribution of w given the data $\mathcal{D} = \{(x_1, y_1), \ldots, (y_n, y_n)\}$.

   (a) Given a new x-value you want to forecast y to minimize the expected square loss. That is, we want to find

   $$\hat{y} = \arg\min_y \mathbb{E}_{y'}(y - y')^2,$$

   where y' has the predictive distribution given x and $\mathcal{D}$. What is $\hat{y}$, and what is the associated expected loss $\mathbb{E}_{y'}(\hat{y} - y')^2$?

   (b) What types of values for σ, Σ_0, n will lead to the prior exerting a lot of influence on our prediction?

   (c) We saw that the Bayesian approach to Gaussian linear regression corresponds to ridge regression. What values in the Bayesian approach correspond to a large amount of regularization?

5. Suppose you are using Bayesian techniques to fit a Poisson regression model. Conditional on x, w, we have $y \sim \text{Pois}(e^{w^T x})$. A colleague, working with his own data set and prior, has given you a function f that returns i.i.d. samples from his posterior distribution on w. Give pseudocode that, given x, lets you sample from the predictive distribution of y given x.