

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 $\ell : \mathcal{A} \times \mathcal{Y} \rightarrow \mathbb{R}$ 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.

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. (★) 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x, y) = x^2 + 4xy + 3y^2$. Compute the gradient $\nabla f(x, y)$.
4. Compute the gradient of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $i = 1, \dots, n$.

Linear Algebra Exercises

- 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 $y \in \mathbb{R}^m$.
 - If $\text{rank}(A) = n$ then solve the normal equations for x .
 - (★) What if $\text{rank}(A) \neq n$?
- Prove that $A^T A + \lambda \mathbf{I}_{n \times n}$ is invertible if $\lambda > 0$ and $A \in \mathbb{R}^{n \times n}$.
- (★) Describe the following set geometrically:

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

Week 2 Pre-Lecture: Concept Check Exercises

Optimization Prerequisites for Lasso

- 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 $x^+ = (x_1^+, \dots, x_d^+)$ 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 $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{R}^d$.]

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

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

Solve the following questions.

- If x in the first problem satisfies $f(x) \in S$ show how to quickly compute (a, b) for the second problem with $a + b = |x|$ and $f(a - b) \in S$.
- 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 $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $S \subseteq \mathbb{R}$ 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.]

Week 2 Lecture: Concept Check Exercises

Starred problems are optional.

Excess Risk Decomposition

1. Let $\mathcal{X} = \mathcal{Y} = \{1, 2, \dots, 10\}$, $\mathcal{A} = \{1, \dots, 10, 11\}$ and suppose the data distribution has marginal distribution $X \sim \text{Unif}\{1, \dots, 10\}$. 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. (★) 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 (in 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.

L_1 and L_2 Regularization

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. (★) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a differentiable function. Prove that $\|\nabla f(x)\|_2 \leq L$ if and only if f is Lipschitz with constant L .
3. (★) Let \hat{w} denote the minimizer for

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

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 $\|\hat{\beta}\|/\|\tilde{\beta}\|$ is always between 0 and 1. Here $\hat{\beta}$ is the parameter vector of the linear model resulting from the regularized least squares problem. Analogously, $\tilde{\beta}$ 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. (★) If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and differentiable at x , the $\partial f(x) = \{\nabla f(x)\}$.
2. Fix $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $x \in \mathbb{R}^n$. Then the subdifferential $\partial f(x)$ is a convex set.
3. (a) True or False: A subgradient of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at x is normal to a hyperplane that globally underestimates 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 $f : \mathbb{R} \rightarrow \mathbb{R}$, if $1, -1 \in \partial f(x)$ then x is a global minimizer of f .
(d) True or False: Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by $f(x_1, x_2) = |x_1| + 2|x_2|$. Compute $\partial f(x_1, x_2)$ for each $x_1, x_2 \in \mathbb{R}^2$.

Week 4 Lecture: Concept Check Exercises

Convexity

1. If $A, B \subseteq \mathbb{R}^n$ are convex, then $A \cap B$ is convex.
2. Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex. Show that $af + bg$ is convex if $a, b \geq 0$.
3. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and differentiable. Prove that if $\nabla f(x) = 0$ then x is a global minimizer.
4. Prove that if $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is strictly convex and x is a global minimizer, then it is the unique global minimizer.
5. Prove that any affine function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is both convex and concave.
6. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex and let $g : \mathbb{R}^m \rightarrow \mathbb{R}^n$ be affine. Then $f \circ g$ is convex.
7. (★★)
 - (a) Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be convex. Show that f has one-sided left and right derivatives at every point.
 - (b) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be convex. Show that f has one-sided directional derivatives at every point.

- (c) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ 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 $x_1, \dots, x_n \in \mathbb{R}^d$ 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 $x_1, \dots, x_n \in \mathbb{R}^d$ and $y_1, \dots, y_n \in \{-1, 1\}$. 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

1. Fix $n > 0$. For $x, y \in \{1, 2, \dots, n\}$ define $k(x, y) = \min(x, y)$. Give an explicit feature map $\varphi : \{1, 2, \dots, n\}$ 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

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

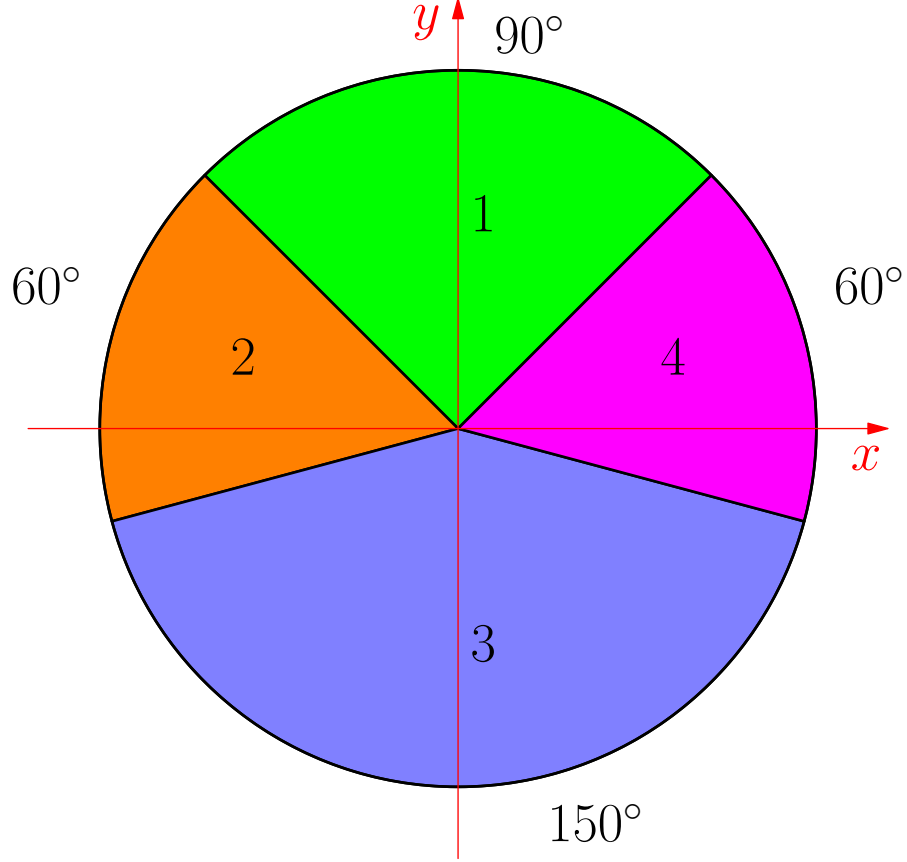
Assume we have a positive semidefinite kernel k .

- (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 $x_1, x_2, \dots, x_n \in \mathbb{R}^n$ and labels $y_1, \dots, y_n \in \{-1, +1\}$. 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 $f_\alpha(x) \rightarrow 0$ as $\|x\|_2 \rightarrow \infty$.
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)?

Week 7 Lecture: Concept Check Exercises

Multiclass

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

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

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

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

Consider the decision function $f \in \mathcal{F}_1$ with w_1, w_2, w_3, w_4 set to the angle bisectors of the corresponding regions. Give $R(f)$.

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

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

where $w \in \mathbb{R}^D$ and $\Psi : \mathbb{R}^2 \times \{1, 2, 3, 4\} \rightarrow \mathbb{R}^D$. 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

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

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 $\mathcal{Y} = \{1, \dots, k\}$, $\mathcal{X} = \mathbb{R}^d$, $w \in \mathbb{R}^D$ and $\psi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^D$. Give a kernelized version of the objective.

Week 9 Lecture: Concept Check Exercises

Trees

- (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.]
- 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.

Bagging

1. Let X_1, \dots, 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, \dots, 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 sample median.

Boosting

1. (★) Show the exponential margin loss is a convex upper bound for the 0 – 1 loss.
2. Show how to perform gradient boosting with the hinge loss.
3. Suppose we are using gradient boosting. On each step we can do a better job of fitting the pseudoresiduals if we allow for deeper trees. Why might deep trees be discouraged while gradient boosting?