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.

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 $h > 0$.

   Solution. The directional derivative is given by

   $$f'(x; u) = \lim_{h\to 0} \frac{f(x + hu) - f(x)}{h} < 0.$$

By the definition of a limit, there must be a $\delta > 0$ such that

$$\frac{f(x+hu) - f(x)}{h} < 0$$

whenever $|h| < \delta$. If we restrict $0 < h < \delta$ then we have

$$f(x+hu) - f(x) < 0 \implies f(x+hu) < f(x)$$

as required.

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

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

Solution. By Cauchy-Schwarz we have, for $\|u\|_2 = 1$,

$$|f'(x;u)| = |\nabla f(x)^T u| \leq \|\nabla f(x)\|_2 \|u\|_2 = \|\nabla f(x)\|_2.$$

Note that

$$\nabla f(x)^T \frac{\nabla f(x)}{\|\nabla f(x)\|_2} = \|\nabla f(x)\|_2 \quad \text{and} \quad \nabla f(x)^T \frac{-\nabla f(x)}{\|\nabla f(x)\|_2} = -\|\nabla f(x)\|_2,$$

so these achieve the maximum and minimum bounds given by Cauchy-Schwarz.

One way to understand the Cauchy-Schwarz inequality is to recall that the dot-product between two vectors $v, w \in \mathbb{R}^d$ can be written as

$$v^T w = \|v\|_2 \|w\|_2 \cos(\theta),$$

where θ is the angle between v and w. This value is maximized at $\cos(0) = 1$ and minimized at $\cos(\pi) = -1$.

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 $\nabla f(x, y)$.

   Solution. Computing the partial derivatives gives

   $$\partial_1 f(x, y) = 2x + 4y \quad \text{and} \quad \partial_2 f(x, y) = 4x + 6y.$$

   Thus the gradient is given by

   $$\nabla f(x, y) = \begin{pmatrix} 2x + 4y \\ 4x + 6y \end{pmatrix}.$$

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.

   Solution. Here we show two methods. In either case we can obtain differentiability by noticing the partial derivatives are continuous.

   (a) Since

   $$f(x) = x^T Ax = \sum_{i,j=1}^{n} a_{ij} x_i x_j$$

   we have

   $$\partial_k f(x) = \sum_{j=1}^{n} (a_{kj} + a_{jk}) x_j$$

   so

   $$\nabla f(x) = (A + A^T)x.$$

   (b) Note that

   $$\begin{aligned} f(x + tv) &= (x + tv)^T A(x + tv) \\ &= x^T Ax + tx^T Av + tv^T Ax + t^2 v^T Av \\ &= f(x) + t(x^T A + x^T A^T)v + t^2(v^T Av). \end{aligned}$$

   Thus

   $$f'(x; v) = \lim_{t\to 0} \frac{f(x + tv) - f(x)}{t} = \lim_{t\to 0} (x^T A + x^T A^T)v + t(v^T Av) = (x^T A + x^T A^T)v.$$

   This shows

   $$\nabla f(x) = (A + A^T)x.$$

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 $A \in \mathbb{R}^{n\times n}$.

Solution. First consider the linear function $g(x) = c^T x$. Note that

$$g(x + tv) = c^T(x + tv) = c^T x + tc^T v \implies \nabla f(x) = c.$$

As the derivative is linear we can combine this with the previous problem to obtain

$$\nabla f(x) = c + (A + A^T)x.$$

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.

   Solution. We give two methods.

   (a) Let $g(x) = x^T Ax$ and $h(x) = x - s$ so that $f(x) = g(h(x))$. By the vector-valued form of the chain rule we have

   $$\nabla f(x) = \nabla g(h(x))^T Dh(x) = (A + A^T)(x - s),$$

   where $Dh(x) = \mathbf{I}_{n \times n}$ is the Jacobian matrix of h.

   (b) We have

   $$(x - s)^T A(x - s) = x^T Ax - s^T(A + A^T)x + s^T As.$$

   Computing the gradient gives

   $$\nabla f(x) = (A + A^T)x - (A + A^T)s = (A + A^T)(x - s).$$

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

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

   Solution.

   (a) We can express $f(w)$ as

   $$f(w) = (Aw - y)^T(Aw - y) + \lambda w^T w = w^T A^T Aw - 2y^T Aw + y^T y + \lambda w^T w.$$

   Applying our previous results gives (noting $w^T w = w^T \mathbf{I}_{n \times n} w$)

   $$\nabla f(w) = 2A^T Aw - 2A^T y + 2\lambda w = 2(A^T A + \lambda \mathbf{I}_{n \times n})w - 2A^T y.$$

(b) Let

$$B = \begin{pmatrix} A \\ \sqrt{\lambda}\mathbf{I}_{n\times n} \end{pmatrix} \quad \text{and} \quad z = \begin{pmatrix} y \\ \mathbf{0}_{n\times 1} \end{pmatrix}$$

written in block-matrix form. Note B is full rank.

(c) The argmin is $w = (A^TA + \lambda\mathbf{I}_{n\times n})^{-1}A^Ty$. To see why the inverse is valid, see the linear algebra questions below.

6. Compute the gradient of

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

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

Solution. As the derivative is linear, we can compute the gradient of each term separately and obtain

$$\nabla f(\theta) = 2\lambda\theta - \sum_{i=1}^{n} \frac{\exp(-y_i\theta^Tx_i)}{1 + \exp(-y_i\theta^Tx_i)} y_ix_i,$$

where we used the techniques from Recitation 1 to differentiate the log terms.

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

Solution.

(a) $a = a^+ - a^-$

(b) $|a| = a^+ + a^-$

(c) $x = x^+ - x^-$

(d) $\|x\|_1 = \mathbf{1}^T(x^+ + x^-)$

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 $p_1^* = p_2^*$.

Solution.

(a) Let $a = x^+$ and $b = x^-$. Then $a + b = |x|$ and $a - b = x$.

(b) Let $x = a - b$ and note that $|x| = |a - b| \leq |a| + |b| = a + b$.

(c) Part a) shows $p_2^* \leq p_1^*$ by letting $\hat{a} = x^+$ and $\hat{b} = x^-$. Part b) shows $p_1^* \leq p_2^*$ by letting $\hat{x} = a - b$.

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.]

Solution. Consider the minimization problem

$$\begin{array}{ll} \text{minimize}_{a,b\in\mathbb{R}^d} & \mathbf{1}^T(a+b) \\ \text{subject to} & f(a-b) \in S, \\ & a_i, b_i \geq 0 \quad \text{for } i = 1, \ldots, d. \end{array}$$

Let p_1^* be the minimum for the original problem, and p_2^* the minimum for our new problem. We first show $p_1^* = p_2^*$. Suppose x is a minimizer for the original problem

and let $a = x^+$ and $b = x^-$. Then by the first question $\mathbf{1}^T(a+b) = \|x\|_1$ and $a - b = x$. This shows $p_2^* \leq p_1^*$. Next suppose (a, b) is a minimizer for our new problem, and let $x = a - b$. Then

$$\|x\|_1 = \|a - b\|_1 = \sum_{i=1}^{d} |a_i - b_i| \leq \sum_{i=1}^{d} |a_i| + |b_i| = \sum_{i=1}^{d} a_i + b_i = \mathbf{1}^T(a + b).$$

This proves $p_1^* \leq p_2^*$.

Finally, given a minimizer (a, b) for the new problem we recover a minimizer x for the original problem by letting $x = a - b$.

Ellipsoids

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

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

Solution. The set is an ellipse with semi-axis lengths $2/\sqrt{3}$ and 2 rotated counter-clockwise by $\pi/4$. Letting $v = (x, y)^T$ and multiplying all terms we get

$$2x^2 + 2xy + 2y^2 = 4.$$

From precalculus we can see this is a conic section, and must be an ellipse or a hyperbola, but more work is needed to determine which one. Instead of proceeding along these lines, let's use linear algebra to give a cleaner treatment that extends to higher dimensions.

Let $A = \begin{pmatrix} 2 & 2 \\ 0 & 2 \end{pmatrix}$. Since $v^T A v$ is a number, we must have $(v^T A v)^T = v^T A^T v$. This gives

$$v^T A^T v = v^T A v = \frac{1}{2} v^T (A^T + A) v = v^T \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} v.$$

Our new matrix is symmetric, and thus allows us to apply the spectral theorem to diagonalize it with an orthonormal basis of eigenvectors. In other words, by rotating our axes we can get a diagonal matrix. Either doing this by hand, or using a computer (Matlab, Mathematica, Numpy) we obtain

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} = Q \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} Q^T \quad \text{where} \quad Q = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix}.$$

The set

$$\left\{ w \in \mathbb{R}^2 \;\middle|\; w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \right\}$$

is an ellipse with semi-axis lengths $2/\sqrt{3}$ and 2 since it corresponds to the equation $3w_1^2 + w_2^2 = 4$. Since Q performs a counter-clockwise rotation by $\pi/4$ we obtain the answer. More concretely,

$$w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \iff (Qw)^T Q \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} Q^T (Qw) = 4 \iff (Qw)^T \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} (Qw) = 4$$

so

$$\{v \mid v^T A v = 4\} = \left\{ Qw \;\middle|\; w^T \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} w = 4 \right\}.$$

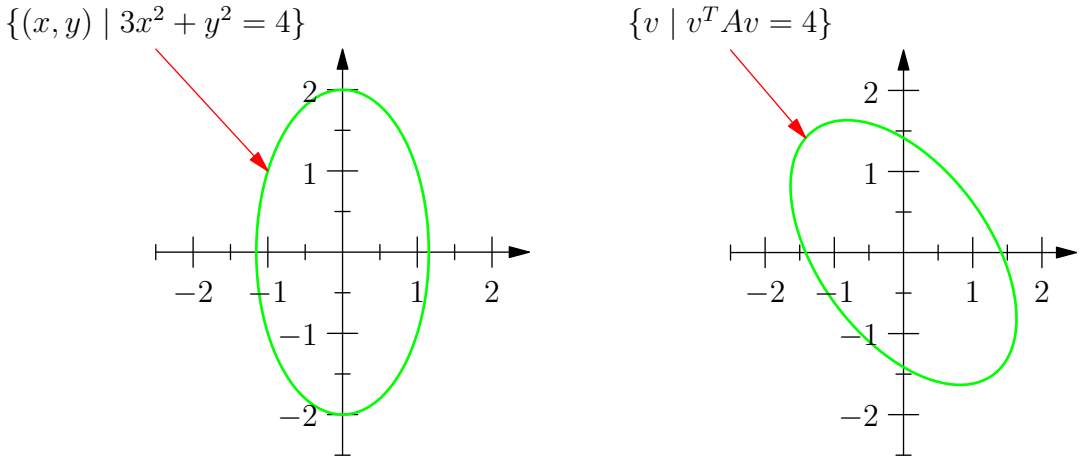


Figure 1: Rotated Ellipse

More generally, the solution to $v^T A v = c$ for $v \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times n}$ and $c > 0$ will be an ellipsoid if A is positive definite. The ith semi-axis will have length $\sqrt{c/\lambda_i}$ where λ_i is the ith eigenvalue of A.

(⋆) 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) (⋆) What if $\mathbf{rank}(A) \neq n$?

   Solution.

   (a) We first show that $\mathbf{rank}(A^T A) = n$ to show that we can invert $A^T A$. By the rank-nullity theorem, we can do this by showing $A^T A$ has trivial nullspace. Note that for any $x \in \mathbb{R}^n$ we have

   $$A^T A x = 0 \implies x^T A^T A x = 0 \implies \|Ax\|_2^2 = 0 \implies Ax = 0 \implies x = 0.$$

This last implication follows since $\mathbf{rank}(A) = n$ so A has trivial nullspace (again by rank-nullity). This proves A^TA has a trivial nullspace, and thus A^TA is invertible. Applying the inverse we obtain

$$x = (A^TA)^{-1}A^Ty.$$

Since A^TA is invertible, our answer for x is unique.

(b) We will show that the equation always has infinitely many solutions x. First note that $\mathbf{rank}(A) \neq n$ implies $\mathbf{rank}(A) < n$ since you cannot have larger rank than the number of columns. Next, recall $\mathbf{rank}(A) = \mathbf{rank}(A^TA)$. Hence, by rank-nullity, A^TA has a non-trivial nullspace, which in turn implies that if there is a solution, there must be infinitely many solutions.

Next note A^T and A^TA have the same column space. To see this, first note that every vector of the form A^TAx must be a linear combination of the columns of A^T, and thus lies in the column space of A^T. Since $\mathbf{rank}(A^TA) = \mathbf{rank}(A) = \mathbf{rank}(A^T)$, this implies A^T and A^TA have the same column spaces.

A specific solution can be computed as $x = (A^TA)^+A^Ty$, where $(A^TA)^+$ is the *pseudoinverse* of A^TA. Of the infinitely many possible solutions x, this gives the one that minimizes $\|x\|_2$. More precisely, $x = (A^TA)^+A^Ty$ solves the optimization problem

$$\begin{array}{ll} \text{minimize} & \|x\|_2 \\ \text{subject to} & A^TAx = A^Ty. \end{array}$$

2. Prove that $A^TA + \lambda\mathbf{I}_{n\times n}$ is invertible if $\lambda > 0$ and $A \in \mathbb{R}^{n\times n}$.

Solution. If $(A^TA + \lambda\mathbf{I}_{n\times n})x = 0$ then

$$0 = x^T(A^TA + \lambda\mathbf{I}_{n\times n})x = \|Ax\|_2^2 + \lambda\|x\|_2^2 \implies x = 0.$$

Thus $A^TA + \lambda\mathbf{I}_{n\times n}$ has trivial nullspace. Alternatively, we could notice that A^TA is positive semidefinite, so adding $\lambda\mathbf{I}_{n\times n}$ will give a matrix whose eigenvalues are all at least $\lambda > 0$. A square matrix is invertible iff its eigenvalues are all non-zero.

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 $R(\hat{f}) - R(f_{\mathcal{F}})$.

   Solution.

   (a) The best decision function is $f^*(x) = x$. The associated risk is 0.

   (b) The best constant function is $f(x) = \mathbb{E}[Y] = \mathbb{E}[X] = 5.5$. This has risk

   $$\mathbb{E}[(Y - 5.5)^2] = \text{Var}(Y) = \frac{33}{4},$$

   by using (or deriving) the formula for the variance of a discrete uniform distribution. Thus the approximation error is 33/4.

   (c) i. The Bayes decision function is affine, so the approximation error is 0.

   ii. The risk is

   $$R(\hat{f}) = \mathbb{E}[(Y - \hat{f}(X))^2] = \mathbb{E}[(X - (X + 1))^2] = 1.$$

   Thus the answer is 1.

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) ?

Solution. Throughout we use the fact that $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

(a) The best decision function is $f(x) = \mathbb{E}[Y|X = x] = a + bx$. This has risk

$$\mathbb{E}[(Y - a - bX)^2] = \mathbb{E}[\mathbb{E}[(Y - a - bX)^2|X]] = \mathbb{E}[1] = 1.$$

(b) The best constant function is given by $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]] = a + b\mathbb{E}[X] = a$. This has risk

$$\mathbb{E}[(Y - a)^2] = \mathbb{E}[\mathbb{E}[(Y - a)^2|X]] = \mathbb{E}[1 + b^2X^2] = 1 + b^2\mathbb{E}[X^2],$$

where

$$\mathbb{E}[X^2] = \int_{-10}^{10} \frac{x^2}{20}\, dx = \frac{2000}{3 \cdot 20} = \frac{100}{3}.$$

Thus the approximation error is $100b^2/3$.

(c) i. There is an affine Bayes decision function, so the approximation error is 0.

ii. Note that

$$\begin{aligned} R(\hat{f}_n) &= \mathbb{E}[(Y - c - dX)^2] = \mathbb{E}[\mathbb{E}[(Y - c - dX)^2|X]] \\ &= \mathbb{E}[1 + ((a - c) + (b - d)X)^2] = 1 + (a - c)^2 + 100(b - d)^2/3. \end{aligned}$$

Thus the estimation error is $(a - c)^2 + 100(b - d)^2/3$.

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.

Solution.

(a) High estimation error due to insufficient data relative to the complexity of your hypothesis space. Can be accompanied by low approximation error indicating a complex hypothesis space.

(b) High approximation error due to an overly simplistic hypothesis space. Can be accompanied by low estimation estimation error due to the large amount of data relative to the (low) complexity of the hypothesis space.

(c) Increased optimization error.

(d) High estimation error.

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.

Solution.

(a) The quantity is random when we consider the training data as a random sample of size n. If we focus on a fixed set of training data then the quantity is deterministic.

(b) False. Note that $\hat{f}_n$ depends on which hypothesis space you have chosen. As an example, imagine having an affine Bayes decision function, and changing the hypothesis space from the set of affine functions to the set of all decision functions. This can cause empirical risk minimization to overfit the training data thus creating a sharp rise in $R(\hat{f}_n) - R(f^*)$.

(c) False, approximation error is a deterministic quantity.

(d) False. The empirical risk of the ERM will often be biased low. This is why we use a test set to approximate its true risk. The issue is that $\hat{f}_n$ depends on the training data so

$$\mathbb{E}\ell(\hat{f}_n(x_i), y_i) \neq \mathbb{E}\ell(\hat{f}_n(x), y)$$

where x, y is a new random draw from the data distribution that isn't in the training data.

(e) i. The space of training sets (i.e., samples of size n from the data generating distribution).

ii. The space of training sets (i.e., samples of size n from the data generating distribution).

iii. The space of all minibatches chosen from the full training set (i.e., samples of of the batch size from the empirical distribution on the full training set).

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.

Solution.

(a) Roughly speaking, more data is better, so we would tend to expect that f_2 will have lower estimation error. That said, this is not always the case, so the relationship cannot be determined.

(b) The approximation error of f_1 will be larger.

(c) The empirical risk of f_1 will be larger.

(d) Roughly speaking, increasing the hypothesis space should increase the estimation error since the approximation error will decrease, and we expect to need more data. That said, this is not always the case, so the answer is the relationship cannot be determined.

(e) Cannot be determined.

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.

   Solution.

   (a) Since the output space is discrete and we are using the $0-1$ loss, our best prediction is the highest probability output conditional on the input. By choosing orange below the line $y = x$ and blue above, we obtain a .1 probability of error. For the $0-1$ loss, probability of error gives the risk.

   (b) No. Any decision tree in $\mathcal{F}$ has finite depth, and thus will divide $[0,1]^2$ into a finite number of rectangles. Thus we cannot produce the decision boundary $y = x$ used by the Bayes decision function.

   (c) Pseudocode follows:

      i. Initialize L to be an empty list of risks.

      ii. Repeat the following M times for some sufficiently large M:

         A. Draw a random sample $(x_1, y_1), \ldots, (x_n, y_n)$ from the data generating distribution.

         B. Obtain a decision function $\tilde{f}$ by running our training algorithm on the generated sample.

         C. Draw a new random sample $(x'_1, y'_1), \ldots, (x'_S, y'_S)$ of size S where S is sufficiently large.

         D. Compute $e = |\{i \mid \tilde{f}(x'_i) \neq y'_i\}|$. That is, the number of times $\tilde{f}$ is incorrect on our new sample.

         E. Add e/S to the list L.

      iii. Compute the sample average and standard deviation of the values in L. Above .176 would be the average and .004 would be the standard deviation.

      Instead of drawing the sample of size S we could have computed the risk analytically.

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.

Solution. Let $C = 1/\lambda$. Then the two objectives differ by a constant factor.

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.

Solution. First suppose $\|\nabla f(x)\|_2 \leq L$ for some $L \geq 0$ and all $x \in \mathbb{R}^n$. By the mean value theorem we have, for any $x, y \in \mathbb{R}^n$,

$$f(y) - f(x) = \nabla f(x + \xi(y - x))^T (y - x),$$

where ξ is some value between 0 and 1. Taking absolute values on each side we have

$$|f(y) - f(x)| = |\nabla f(x + \xi(y - x))^T (y - x)| \leq \|\nabla f(x + \xi(y - x))\|_2 \|y - x\|_2$$

by Cauchy-Schwarz. Applying our bound on the gradient norm proves f is Lipschitz with constant L.

Conversely, suppose f is Lipschitz with constant L. Note that

$$|\nabla f(x)^T v| = |f'(x;v)| = \left|\lim_{t\to 0} \frac{f(x+tv)-f(x)}{t}\right| \leq \lim_{t\to 0} \frac{|t|L\|v\|}{|t|} = L\|v\|.$$

Letting $v = \nabla f(x)$ we obtain $\|\nabla f(x)\|_2^2 \leq L\|\nabla f(x)\|_2$ giving the result.

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.

Solution. Note that $\|w\|_2 \leq \|w\|_1 \leq r$, so the argument from class gives the result. To see the inequality, note that

$$\|w\|_1^2 = (|w_1| + \cdots + |w_n|)^2 \geq |w_1|^2 + \cdots + |w_n|^2 = \|w\|_2^2.$$

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 $[0,1]$?

Solution. We assume Ivanov regularization (since Tikhonov is equivalent). We know that

$$\frac{1}{n}\sum_{i=1}^n (\tilde{w}^T x_i - y_i)^2 \leq \frac{1}{n}\sum_{i=1}^n (\hat{w}^T x_i - y_i)^2$$

since $\tilde{w}$ is the solution to the unconstrained minimization. But if $\|\tilde{w}\| \leq \|\hat{w}\|$ then $\|\tilde{w}\|$ is feasible for the regularized problem, so $\|\hat{w}\| = \|\tilde{w}\|$. Thus $\|\tilde{w}\| \geq \|\hat{w}\|$.

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

Solution. Suppose you have a model $y = w^T x$ where x_1 is a very correlated with y, but the feature is measured in meters. Thus $w_1 = 4$ would mean each increase in x_1 by 1 meter yields an increase in y by 4. Now suppose we change the units of w_1 to kilometers by scaling it. This would require us to change w_1 to 4000 to achieve the same decision function. While this has no effect on the loss $(y - w^T x)^2$ it has a significant effect on $\lambda\|w\|_2^2$ or $\lambda\|w\|_1$. For example, even if $x_2, \ldots, x_n$ had very little relationship with y, we would still undervalue w_1 due to the 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 $\partial f(x) = \{\nabla f(x)\}$.

   Solution. By the gradient (first-order) conditions for convexity, we know that $\nabla f(x) \in \partial f(x)$. Next suppose $g \in \partial f(x)$. This means that for all $v \in \mathbb{R}^n$ and $h \in \mathbb{R}$ we have

   $$f(x + hv) \geq f(x) + hg^T v \implies \frac{f(x + hv) - f(x)}{h} \geq g^T v.$$

   Using $-h$ in place of h gives

   $$f(x - hv) \geq f(x) - hg^T v \implies g^T v \geq \frac{f(x - hv) - f(x)}{-h}.$$

   Taking limits as $h \to 0$ gives

   $$\nabla f(x)^T v \geq g^T v \geq \nabla f(x)^T v.$$

   Thus all terms are equal. Subtracting gives

   $$(\nabla f(x) - g)^T v = 0,$$

   which holds for all $v \in \mathbb{R}^n$. Letting $v = \nabla f(x) - g$ proves

   $$\|\nabla f(x) - g\|_2^2 = 0$$

   giving the result.

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.

   Solution. Let $g_1, g_2 \in \partial f(x)$ and $t \in (0, 1)$. We must show $(1 - t)g_1 + tg_2$ is a subgradient. Note that, for any $y \in \mathbb{R}^n$, we have

   $$\begin{aligned} f(x) + ((1 - t)g_1 + tg_2)^T (y - x) &= (1 - t)(f(x) + g_1^T (y - x)) + t(f(x) + g_2^T (y - x)) \\ &\leq (1 - t)f(y) + tf(y) \\ &= f(y). \end{aligned}$$

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.

Solution.

(a) False. The underestimating hyperplane is a subset of $\mathbb{R}^{n+1}$ but a subgradient is an element of $\mathbb{R}^n$.

(b) False. In lab we considered $f(x_1, x_2) = |x_1| + 2|x_2|$ and noted that $(1, -2) \in \partial f(3, 0)$ but $(-1, 2)$ is not a descent direction.

(c) True. The subdifferential of f at x is convex, and thus contains 0. If 0 is a subgradient of f at x, then x is a global minimizer.

(d) False. Suppose $f : \mathbb{R} \to \mathbb{R}$ is defined by $f(x) = x^2$. Then $\partial f(1) = \{2\}$, and thus doesn't contain 2α for $\alpha \in [0, 1)$.

(e) False. A counterexample is $f(x) = -e^{-x^2}$. The converse is true though. Functions that have convex sublevel sets are called *quasiconvex*.

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 $x_1, x_2 \in \mathbb{R}^2$.

Solution. Write $f(x_1, x_2) = f_1(x_1, x_2) + f_2(x_1, x_2)$ where $f_1(x_1, x_2) = |x_1|$ and $f_2(x_1, x_2) = 2|x_2|$. When $x_1 \neq 0$ we have $\partial f_1(x_1, x_2) = \{(\operatorname{sgn}(x_1), 0)^T\}$ and when $x_1 = 0$ we have

$$\partial f_1(x_1, x_2) = \{(b, 0)^T \mid b \in [-1, 1]\}.$$

When $x_2 \neq 0$ we have $\partial f_2(x_1, x_2) = \{(0, 2\operatorname{sgn}(x_2))^T\}$ and when $x_2 = 0$ we have

$$\partial f_2(x_1, x_2) = \{(0, c)^T \mid c \in [-2, 2]\}.$$

Combining we have

$$\partial f(x_1, x_2) = \partial f_1(x_1, x_2) + \partial f_2(x_1, x_2),$$

where we are summing sets. Recall that if $A, B \subseteq \mathbb{R}^n$ then

$$A + B = \{a + b \mid a \in A, b \in B\}.$$

This gives 4 cases:

(a) If $x_1, x_2 \neq 0$ this gives $\partial f(x_1, x_2) = \{(\operatorname{sgn}(x_1), 2\operatorname{sgn}(x_2))^T\}$.

(b) If $x_1 = 0$ and $x_2 \neq 0$ we have $\partial f(x_1, x_2) = \{(b, 2\operatorname{sgn}(x_2))^T \mid b \in [-1, 1]\}$.

(c) If $x_1 \neq 0$ and $x_2 = 0$ we have $\partial f(x_1, x_2) = \{(\operatorname{sgn}(x_1), c)^T \mid c \in [-2, 2]\}$.

(d) If $x_1 = 0$ and $x_2 = 0$ we have $\partial f(x_1, x_2) = \{(b, c)^T \mid b \in [-1, 1], c \in [-2, 2]\}$.

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.

   Solution. Let $x, y \in A \cap B$ and $t \in (0,1)$. Since A, B are convex, we have

   $$(1-t)x + ty \in A \quad \text{and} \quad (1-t)x + ty \in B.$$

   Thus $(1-t)x + ty \in A \cap B$.

2. Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be convex. Show that $af + bg$ is convex if $a, b \geq 0$.

   Solution. Let $x, y \in \mathbb{R}^n$ and $\theta \in (0,1)$. Then

   $$\begin{aligned}(af+bg)((1-\theta)x+\theta y) &= af((1-\theta)x+\theta y) + bg((1-\theta)x+\theta y) \\ &\leq a[(1-\theta)f(x)+\theta f(y)] + b[(1-\theta)g(x)+\theta g(y)] \\ &= (1-\theta)(af+bg)(x) + \theta(af+bg)(y).\end{aligned}$$

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.

   Solution. Suppose $\nabla f(x) = 0$. The gradient (or first-order) characterization of convexity says

   $$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

   for all y. If $\nabla f(x) = 0$ then this says $f(y) \geq f(x)$ for all x.

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.

   Solution. Suppose y is also a global minimizer with $y \neq x$. Then

   $$f((y+x)/2) < f(y)/2 + f(x)/2 = f(x)$$

   contradicting the fact that $f(x)$ was a global minimizer.

5. Prove that any affine function $f : \mathbb{R}^n \to \mathbb{R}$ is both convex and concave.

   Solution. Recall that f has the form $f(x) = w^T x + b$ where $w \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Then, for $x, y \in \mathbb{R}^n$ and $\theta \in (0,1)$,

   $$f((1-\theta)x+\theta y) = w^T((1-\theta)x+\theta y)+b = (1-\theta)(w^T x+b)+\theta(w^T y+b) = (1-\theta)f(x)+\theta f(y).$$

   This shows f is convex. But the same holds if we replace w with $-w$ and b with $-b$. Hence f is also 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.

   Solution. Write $g(x) = Ax + b$ where $A \in \mathbb{R}^{n\times m}$ and $b \in \mathbb{R}^n$. For $x, y \in \mathbb{R}^m$ and $t \in (0,1)$ we have

   $$\begin{aligned} f(g((1-t)x+ty)) &= f((1-t)(Ax+b)+t(Ay+b)) \\ &\leq (1-t)f(Ax+b)+tf(Ay+b) \\ &= (1-t)f(g(x))+tf(g(y)). \end{aligned}$$

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).

   Solution. We first prove the following lemma.

   Lemma 1. *If $f : \mathbb{R} \to \mathbb{R}$ is convex and $x < y < z$ then*

   $$\frac{f(y)-f(x)}{y-x} \leq \frac{f(z)-f(x)}{z-x}.$$

Proof. Let $t \in (0,1)$ satisfy $(1-t)x + tz = y$. By convexity we have

$$f(y) = f((1-t)x + tz) \leq (1-t)f(x) + tf(z)$$

giving

$$\frac{f(y) - f(x)}{y - x} \leq \frac{(1-t)f(x) + tf(z) - f(x)}{(1-t)x + tz - x} = \frac{t(f(z) - f(x))}{t(z-x)} = \frac{f(z) - f(x)}{z - x}.$$

□

(a) For the right derivative, we will show

$$\lim_{y \downarrow x} \frac{f(y) - f(x)}{y - x} = \inf_{y > x} \frac{f(y) - f(x)}{y - x} =: L.$$

Fix $\epsilon > 0$ and choose $y' > x$ so that

$$\frac{f(y') - f(x)}{y' - x} < L + \epsilon.$$

Letting $\delta = y' - x$, the lemma shows that

$$\frac{f(y) - f(x)}{y - x} < L + \epsilon$$

for any $y < x + \delta$ proving the limit exists.

For the left derivative, we could repeat the above, or note that $g(t) = 2x - t$ is affine, so $f \circ g$ is convex. By the above

$$\lim_{y \downarrow x} \frac{f(g(y)) - f(g(x))}{y - x} = \lim_{y \downarrow x} \frac{f(2x - y) - f(x)}{y - x} = \lim_{h \downarrow 0} \frac{f(x - h) - f(x)}{h}$$

exists, where $h = y - x$. This proves the left derivative exists as well.

(b) Fix $x, v \in \mathbb{R}^n$ and let $g : \mathbb{R} \to \mathbb{R}^n$ be defined by $g(t) = x + tv$. Then $f \circ g$ is convex, and thus the previous part applies. But the right derivative of g at 0 is the one-sided directional derivative of f at x in the direction v:

$$\lim_{h \downarrow 0} \frac{f(g(h)) - f(g(0))}{h} = \lim_{h \downarrow 0} \frac{f(x + hv) - f(x)}{h}.$$

(c) Let y be a minimizer of f and let $g(t) = x + t(y - x)$. By the arguments in the first part above, the value

$$\frac{f(g(1)) - f(g(0))}{1 - 0} = f(y) - f(x) < 0$$

is an upper bound on the right derivative of g at 0. But this is a directional derivative, by the argument in the second part above.

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?

   Solution. Let H_{ij} denote the height of the person in row i and column j. Then

   $$a = \max_j \min_i H_{ij} \leq \min_i \max_j H_{ij} = b,$$

   by the max-min inequality.

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.

   Solution.

   $$\begin{array}{ll} \text{minimize}_{r,c} & r \\ \text{subject to} & \|x_i - c\|_2 \leq r \quad \text{for } i = 1, \ldots, n. \end{array}$$

   This problem is convex since norms are convex, so $f_i(c) = \|x_i - c\|_2$ is convex (composition of convex with affine).

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.

   Solution. Solve the hard-margin SVM problem

   $$\begin{array}{ll} \text{minimize}_{w,b} & \|w\|_2^2 \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 \quad \text{for all } i = 1, \ldots, n. \end{array}$$

   If the resulting problem is feasible, then the data is 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?

   Solution.

(a) $L(x, \lambda) = \|Ax - y\|_2^2 + \lambda(\|x\|_2^2 - r^2)$. Note that this is a shifted version of the Tikhonov objective.

(b)
$$\sup_{\lambda \succeq 0} L(x, \lambda) = \begin{cases} +\infty & \text{if } \|x\|_2^2 > r^2, \\ \|Ax - y\|_2^2 & \text{otherwise.} \end{cases}$$

Note that the original Ivanov minimization is then just

$$\inf_x \sup_{\lambda \succeq 0} L(x, \lambda).$$

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 $k(x, y) = \varphi(x)^T \varphi(y)$.

   Solution. Define $\varphi(x) = (\mathbf{1}(x \leq 1), \mathbf{1}(x \leq 2), \ldots, \mathbf{1}(x \leq n))$. Then $\varphi(x)^T \varphi(y) = \min(x, y)$.

2. Show that $k(x, y) = (x^T y)^4$ is a positive semidefinite kernel on $\mathbb{R}^d \times \mathbb{R}^d$.

   Solution. $k_1(x, y) = x^T y$ is a psd kernel, since $x^T y$ is an inner product on $\mathbb{R}^d$. Using the product rule for psd kernels, we see that

   $$k(x, y) = k_1(x, y)k_1(x, y)k_1(x, y)k_1(x, y) = k_1(x, y)^4$$

   is psd as well.

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.

   Solution. Fix $x_1, \ldots, x_n \in \mathbb{R}^d$ and let X denote the matrix that has x_i^T as its ith row. Then note that $(XAX^T)_{ij} = x_i^T A x_j = k(x_i, x_j)$. Thus we are done if we can show XAX^T is positive semidefinite. But note that, for any $\alpha \in \mathbb{R}^n$,

   $$\alpha^T XAX^T \alpha = (X^T \alpha)^T A (X^T \alpha) \geq 0,$$

   since A is positive semidefinite.

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.

   Solution.

   (a) $J(\alpha) = \|K\alpha - y\|_1 + \lambda \alpha^T K \alpha$, where $K_{ij} = k(x_i, x_j)$. Here x_i^T is the ith row of X.

   (b) $f_\alpha(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$.

5. Show that the standard 2-norm on $\mathbb{R}^n$ satisfies the parallelogram law.

   Solution.

   $$\begin{aligned} \|x - y\|_2^2 + \|x + y\|_2^2 &= (\|x\|_2^2 - 2x^T y + \|y\|_2^2) + (\|x\|_2^2 + 2x^T y + \|y\|_2^2) \\ &= 2\|x\|_2^2 + 2\|y\|_2^2. \end{aligned}$$

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.

   Solution. By selecting σ sufficiently small (say, much smaller than $\min_{i \neq j} \|x_i - x_j\|_2$) we can use $\alpha_i = y_i$ and get very pointy spikes at each data point. [Note: This is not possible if any repeated points have different labels, which is not unusual in real data.]

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 $\|x\|_2 \to \infty$.

   Solution. Since $f_\alpha(x) = \sum_{i=1}^n \alpha_i k(x_i, x)$ we have

$$\lim_{\|x\|_2 \to \infty} f_\alpha(x) = \lim_{\|x\|_2 \to \infty} \sum_{i=1}^n \alpha_i \exp\left(-\frac{\|x_i - x\|_2^2}{2\sigma^2}\right) = \sum_{i=1}^n \alpha_i \lim_{\|x\|_2 \to \infty} \exp\left(-\frac{\|x_i - x\|_2^2}{2\sigma^2}\right) = 0.$$

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)?

   Solution.

   (a) $\{w^* + v \mid v \in \text{null}(X)\}$. Using the standard inner product on $\mathbb{R}^n$, we can also write $\text{null}(X)$ as the set of all vectors orthogonal to the row space of X.

   (b) w^* lies in the row space of X.

Multiclass: Concept Check

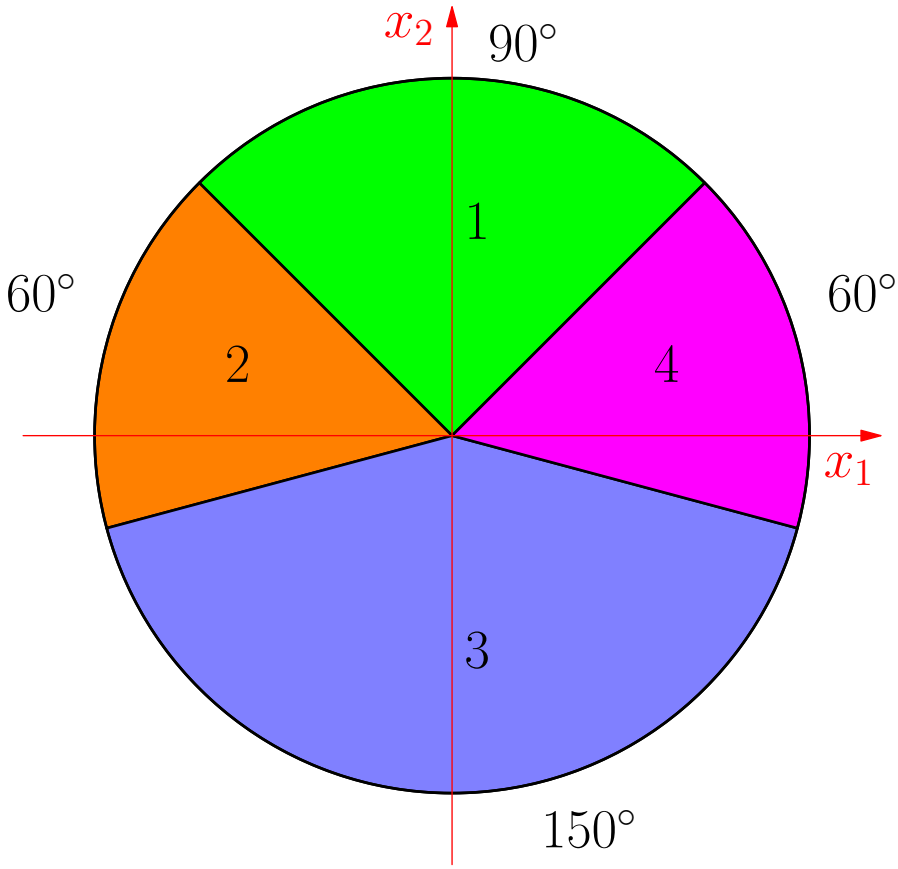
Multiclass Learning Objectives

- 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

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

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

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

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 $\mathcal{F}$:

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

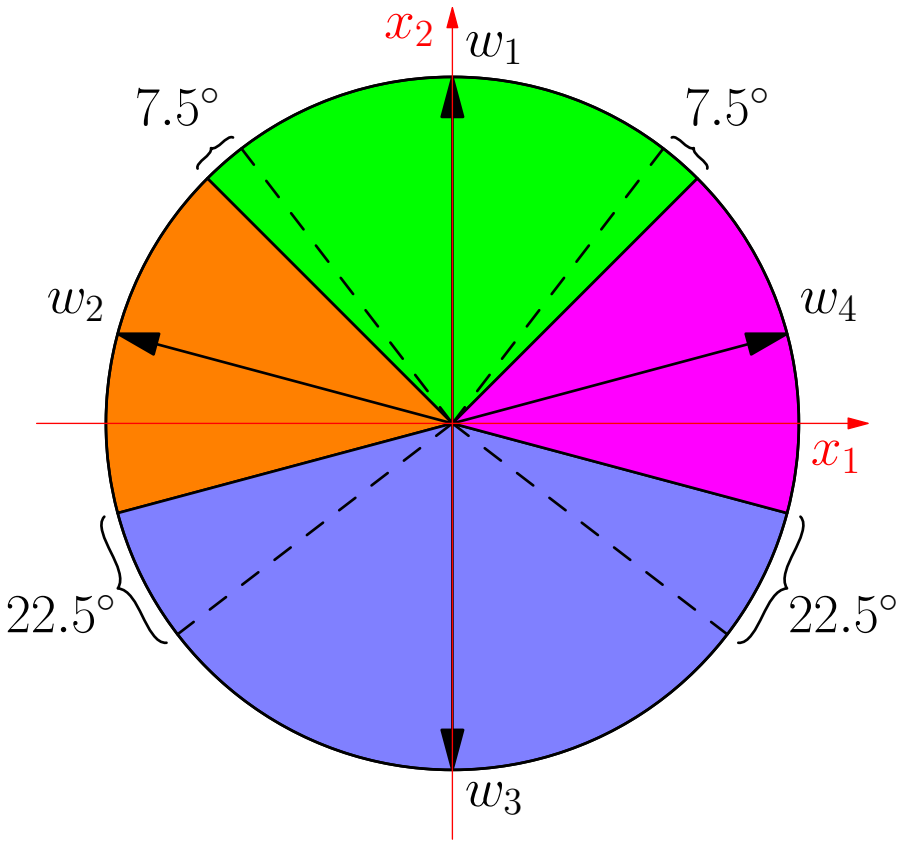
where $w \in \mathbb{R}^D$ and $\Psi : \mathbb{R}^2 \times \{1, 2, 3, 4\} \to \mathbb{R}^D$. Give w, Ψ corresponding to $f_{\mathcal{F}_2}$, the decision function minimizing the risk over $\mathcal{F}_2$.

Solution.

(a) Let $w_1 = (0, 1)^T$, $w_2 = (-1, 0)^T$, $w_3 = (0, -c)^T$, $w_4 = (1, 0)^T$, where $c = \cot \frac{\pi}{12} = 2 + \sqrt{3}$. The corresponding risk is 0. To see how c was computed, consider the boundary between the magenta and blue regions. The division occurs along the vector $(\cos(\pi/12), -\sin(\pi/12))$. Note that

$$w_4^T(\cos(\pi/12), -\sin(\pi/12)) = \cos(\pi/12) = w_3^T(\cos(\pi/12), -\sin(\pi/12)).$$

(b) We have $w_1 = (0, 1)$, $w_3 = (0, -1)$, $w_2 = (-\cos(\pi/2), \sin(\pi/12))$, $w_4 = (\cos(\pi/12), \sin(\pi/12))$. This gives the image below.



The dashed lines above are the boundaries of the 4 regions. The resulting risk is $(7.5 + 7.5 + 22.5 + 22.5)/360 = 1/6$.

(c) Let $w = (0, 1, -1, 0, 0, -\cot(\pi/12), 1, 0) \in \mathbb{R}^8$ and define

$$\psi(x, i) = x_1 e_{2i-1} + x_2 e_{2i} \in \mathbb{R}^8$$

where e_j is the vector with 1 in the jth position and 0 elsewhere.

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.

Solution. Let $\Psi(x, y) = \frac{1}{2}yx$ for $y \in \{-1, +1\}$. Then we have, for $y \neq y_i$,

$$1 - m_{i,y}(w) = 1 - (w^T x_i y_i - w^T x_i y)/2 = \begin{cases} 1 + w^T x_i & \text{if } y_i = -1, \\ 1 - w^T x_i & \text{if } y_i = +1. \end{cases}$$

This gives $1 - m_{i,y}(w) = 1 - y_i w^T \varphi(x_i)$.

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

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

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

Solution.

(a) Evaluate $w^T \psi(x, i)$ for $i = 1, \ldots, k$.

(b) Forecast the value i that gives the largest $w^T \psi(x, i)$ value.

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.

Solution. Let $X \in \mathbb{R}^{nk \times D}$ matrix that has each $\Psi(x_i, y)^T$ as rows for each $i = 1, \ldots, n$ and $y = 1, \ldots, k$. More precisely, $\Psi(x_i, y)^T$ will be in row $(i-1)k + y$ of X. By the representer theorem, a solution, if it exists, must have the form $w^* = X^T \alpha$. Let $XX^T = K$, the Gram matrix. Then we have

$$m_{i,y}(w) = w^T \Psi(x_i, y_i) - w^T \Psi(x_i, y) = (K\alpha)_{(i-1)k+y_i} - (K\alpha)_{(i-1)k+y},$$

and $\|w\|_2^2 = \alpha^T K \alpha$. Substituting we have

$$J(\alpha) = \frac{1}{2}\alpha^T K \alpha + \frac{C}{n}\sum_{i=1}^{n} \max_{y \neq y_i} (1 - ((K\alpha)_{(i-1)k+y_i} - (K\alpha)_{(i-1)k+y}))_+.$$

Note that the Gram matrix K is $nk \times nk$, and thus can be infeasible to store or compute for nk large.

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.]

   Solution.

   (a) Given a fixed tree, if we split a leaf node we add a single leaf to the tree. Thus k splits corresponds to $k+1$ leaves.
   (b) A tree of height k can have at most 2^k regions (leaves).
   (c) A tree of height $\lceil \log_2(n) \rceil$ is sufficient to distinguish all possible values for the first feature. At each leaf we can then put another tree of this height that distinguishes the second feature, and so forth. These give an upper bound of $d\lceil \log_2(n) \rceil$.

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.

   Solution.

```
import numpy as np
def bestSplit(y) :
    '''
    Greedily computes the best splitting point for the current node.
    Assumes there is at least 2 values. There are more numerically
    stable ways of doing this
    [see The Art of Computer Programming p. 232, Vol 2, 3rd Edition]
    @param y : array-like, shape = [n_samples,], contains the
    output values for each sample. Assumes the
    values are sorted by the corresponding first
    feature values.
    @return the value i such that inputs 0,...,i belong in the
    left subtree.
    '''
    sums = np.cumsum(y) #partial sums of y-values
    sumsq = np.cumsum([a**2 for a in y]) #partial sums of squared y-values
    S = sums[-1] #sum of all y-values
    SS = sumsq[-1] # sum of all squared y-values
    bestIdx = -1
    bestLoss = None
    N = len(y)
    for idx in range(N-1) :
        leftLoss = sumsq[idx] - sums[idx]**2/(idx+1.0)
        rightLoss = (SS-sumsq[idx])-(S-sums[idx])**2/(N-(idx+1.0))
        loss = leftLoss+rightLoss
        if bestIdx == -1 or loss < bestLoss :
            bestIdx = idx
            bestLoss = loss
    return bestIdx,bestLoss
```

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.

Solution.

(a) Left: 3/6, Right: 1/3, Total: $6(3/6) + 3(1/3) = 4$

(b) Left: $3/6(3/6)+2/6(4/6)+1/6(5/6) = 22/36$, Right: $1/3(2/3)+2/3(1/3) = 4/9$, Total: $6(22/36)+3(4/9) = 30/6 = 5$

(c) Left: $-3/6\log(3/6)-2/6\log(2/6)-1/6\log(1/6)$, Right: $-1/3\log(1/3)-2/3\log(2/3)$, Total:

$$6[-3/6\log(3/6) - 2/6\log(2/6) - 1/6\log(1/6)] + 3[-1/3\log(1/3) - 2/3\log(2/3)].$$

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?

   Solution. Recall that the sample mean has variance

   $$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\text{Var}(X_1)}{n} = \frac{\sigma^2}{n},$$

   with standard error $\sigma/\sqrt{n}$. Thus we have

   $$\sigma/\sqrt{n} < .01 \iff n > 10000\sigma^2.$$

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.

Solution.

(a) i. Draw B bootstrap samples $D^1, \ldots, D^B$ each of size $2n+1$. The samples are formed by drawing uniformly with replacement from the original data set $X_1, \ldots, X_{2n+1}$. We will make a total of $B(2n+1)$ draws.

ii. For each D^i compute the corresponding median $\hat{m}_i$.

iii. Compute the sample variance of the B medians $m_1, \ldots, m_B$.

(b) i. Draw B bootstrap samples $D^1, \ldots, D^B$ each of size $2n+1$. The samples are formed by drawing uniformly with replacement from the original data set $X_1, \ldots, X_{2n+1}$. We will make a total of $B(2n+1)$ draws.

ii. For each D^i compute the corresponding median $\hat{m}_i$.

iii. Compute the 2.5% and 97.5% sample quantiles of the list $\hat{m}_1, \ldots, \hat{m}_B$. Use these as the estimates of the left and right endpoints of the confidence interval, respectively.

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 $f_X(x) = \lambda e^{-\lambda x}$.

   Solution.

   (a) The likelihood L is given by

   $$L(p; x_1, \ldots, x_n) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p$$

   giving a log-likelihood

   $$\log L(p; x_1, \ldots, x_n) = n \log p + \left(\sum_{i=1}^{n} x_i - 1 \right) \log(1-p).$$

   Differentiating gives

   $$\frac{d}{dp} \log L(p; x_1, \ldots, x_n) = \frac{n}{p} - \frac{\sum_{i=1}^{n} x_i - 1}{1-p}.$$

   Solving for a critical point we get

   $$\frac{d}{dp} \log L(p; x_1, \ldots, x_n) = 0 \iff \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{p} \iff p = \frac{n}{\sum_{i=1}^{n} x_i}.$$

   By the first or second derivative tests, this is the maximum. Thus the answer is

   $$\hat{p}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

   (b) The likelihood L is given by

   $$L(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

   giving a log-likelihood

   $$\log L(\lambda; x_1, \ldots, x_n) = n \log \lambda - \lambda \sum_{i=1}^{n} x_i.$$

Differentiating gives

$$\frac{d}{dp}\log L(p; x_1, \ldots, x_n) = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i.$$

Solving for a critical point we get

$$\frac{d}{dp}\log L(p; x_1, \ldots, x_n) = 0 \iff \lambda = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

By the first or second derivative tests, this is a maximum. Thus the answer is

$$\hat{\lambda}_{\text{MLE}} = \frac{n}{\sum_{i=1}^{n} x_i}.$$

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.

   Solution. The likelihood L is given by

   $$L(w; x_1, y_1, \ldots, x_n, y_n) = \prod_{i=1}^{n} \frac{\mathbf{1}(y_i \le e^{w^T x_i})}{e^{w^T x_i}}.$$

   Taking logs we get

   $$-\sum_{i=1}^{n} w^T x_i = -w^T \left(\sum_{i=1}^{n} x_i\right)$$

   if $y_i \le \exp(w^T x_i)$ for all i, or $-\infty$ otherwise. Thus we obtain the linear program

   $$\begin{aligned} &\text{minimize} && w^T \left(\sum_{i=1}^{n} x_i\right) \\ &\text{subject to} && \log(y_i) \le w^T x_i \quad \text{for } i = 1, \ldots, n. \end{aligned}$$

3. Explain why softmax is related to computing the maximum of a list of values.

   Solution. Let $x_1, \ldots, x_n \in \mathbb{R}$. Let $\text{ArgMax}(x_1, \ldots, x_n)$ denote a 1-hot encoding of the argmax function:

   $$\text{ArgMax}(x_1, \ldots, x_n) = \left(\mathbf{1}(\arg\max_i x_i = 1), \ldots, \mathbf{1}(\arg\max_i x_i = n)\right).$$

   Recall that softmax has the following definition:

   $$\text{softmax}_\lambda(x_1, \ldots, x_n) = \frac{1}{\sum_{i=1}^{n} e^{\lambda x_i}} \left(e^{\lambda x_1}, \ldots, e^{\lambda x_n}\right),$$

where $\lambda > 0$ is a fixed parameter. We claim that softmax is a differentiable approximation to ArgMax. Consider what happens when we let $x_j \to \infty$ while keeping the other values fixed. Then

$$\frac{e^{\lambda x_j}}{\sum_{i=1}^n e^{\lambda x_i}} \to 1$$

and

$$\frac{e^{\lambda x_k}}{\sum_{i=1}^n e^{\lambda x_i}} \to 0$$

for all $k \neq j$. For example, suppose $x_1 = 1$, $x_2 = -3$, $x_3 = 5$. Then

$$\text{softmax}_1(1, -3, 5) = (0.0180, 0.0003, 0.9817)$$

while

$$\text{ArgMax}(1, -3, 5) = (0, 0, 1).$$

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?

Solution. From Bayes' theorem[1], we have:

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ &\propto (\theta^x \exp(-\theta)) \left(\theta^{\alpha-1} \exp(-\beta\theta)\right) \\ &= \theta^{x+\alpha-1} \exp\left(-(\beta+1)\theta\right) \\ &\propto \mathcal{G}(\alpha + x, \beta + 1) \end{aligned}$$

This shows that the gamma is the conjugate prior to the Poisson. Also, note here we exploit a common trick: we manipulate the numerator, ignoring constants independent of θ. If we can recognize the functional form as belonging to a distribution family we know, we can simply identify the parameters and trust that the distribution normalizes!

[1] Actually from Roman Garnett, from whom this problem was taken.

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.

   Solution. Note that both prior distributions are from the Beta family.

   (a) We have

$$\xi_A(\theta|x) \propto f(x|\theta)\xi_A(\theta) \propto \theta^{711}(1-\theta)^{290}$$

   and

$$\xi_B(\theta|x) \propto f(x|\theta)\xi_B(\theta) \propto \theta^{713}(1-\theta)^{290}.$$

Thus the posteriors from A and B are both beta with parameters $(712, 291)$ and $(714, 291)$, respectively.

(b) The respective means are $\frac{712}{1003}$ and $\frac{714}{1005}$.

(c) In general the two means are given by

$$\frac{a+2}{1003} \quad \text{and} \quad \frac{a+4}{1005}.$$

The difference is less than $2/1000 = .002$.

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.

Solution. The frequentist student is totally incorrect, since they have misunderstood what a frequentist confidence interval is. Using frequentist methodology, θ is not a random variable, so it doesn't make sense to say it lies in some fixed interval $[L_F, R_F]$. The correct interpretation is that if independent experiments like this were repeated, then at least 95% of the time $[L_F, R_F]$ will contain θ. That is, the interval is random not θ.

We can say that the Bayesian student is consistent. Recall that to compute the credible set, the Bayesian student had to introduce some prior distribution π on θ. What we can say is if someone believes π is correct, then it is rational, given the data, to conclude that θ will lie in the posterior credible set with probability 95%.

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 $b = 1$.]

Solution. As hinted, let's try $a = 10$, $b = 1$ and 9 coin flips all landing tails. The prior variance is given by

$$\frac{10 \cdot 1}{(10+1)^2(10+1+1)} = \frac{5}{726} \approx .0069$$

while the posterior variance is given by

$$\frac{10 \cdot 10}{(10+10)^2(10+10+1)} = \frac{1}{84} \approx .0119.$$

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?

   Solution.

   (a) We have $\hat{y} = \mu_1^T x$ with expected loss $x^T \Sigma_1 x + \sigma^2$, the mean and variance of the predictive distribution.

   (b) i. High σ meaning low certainty in data.
   ii. Small Σ_0 meaning high certainty in prior. A covariance matrix is small if its eigenvalues are small.
   iii. Small n meaning not a lot of data to learn from.

   (c) Small Σ_0 meaning high certainty in prior.

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.

   Solution. Pseudocode follows:

   (a) Draw w from f.

   (b) Draw y from $\text{Pois}(e^{w^T x})$.

   (c) Return y.