CS 189 Introduction to Machine Learning
Spring 2018 gs1

1 Least Squares

(a) Consider the weighted least squares problem where $x_i \in \mathbb{R}^m$, $y \in \mathbb{R}^n$:

$$\sum_{i=1}^{n} c_i (w^T x_i - y_i)^2 : c_i \geq 0$$

Show that this problem can be written in matrix form where C is a diagonal matrix. What are X and y? What is C?

(b) Now consider adding a normalizing constraint:

$$(Xw - y)^T C (Xw - y) + \lambda \|w\|_2^2$$

Show that this problem is equivalent to:

$$\|\hat{X}w - \hat{y}\|_2^2$$

How would you form $\hat{X}$ and $\hat{y}$?
Hint: What properties of C can be used to simplify this problem?

(c) In homework, we saw how we can think of Ridge Regression as a constrained version of Ordinary Least Squares. To review we can rewrite:

$$\begin{aligned} &\min \|Xw - y\|_2^2 \\ &\text{s.t. } \|w\|_2^2 \leq \beta^2 \end{aligned}$$

As

$$\min \|Xw - y\|_2^2 + \lambda \|w\|_2^2$$

Where λ is a parameter denoting the "price" we pay for violating the constraint.
Now, we will consider a similar constrained optimization problem:

$$\begin{aligned} &\min \|Xw - y\|_2^2 \\ &\text{s.t. } \|w - v\|_2^2 \leq \beta^2 \end{aligned}$$

What does this problem represent in terms of prior belief (informally is fine). Solve the problem, and explain why the solution makes sense.
Hint: There is a long way and an elegant way of solving this
Hint: The elegant way does not require taking a derivative

(d) Regarding kernelized ridge regression, Note 7 makes the claim that:

$$w^* = \Phi^\top(\Phi\Phi^\top + \lambda I)^{-1}y = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top y$$

(and that we prefer to use the more efficient approach depending on the number of samples we have vs. the degree p of the polynomial features we wish to use.)

Prove this claim for $\lambda \neq 0$.

2 MLE/MAP

(a) Consider a biased coin with probability of heads p. Suppose we flip the coin n times to get samples $X_1, X_2, \ldots, X_n$, which we know come from a Bernoulli distribution. Recall that this means for some X_i, the outcome will be heads with probability p and tails with probability $1-p$. Define the likelihood function $\mathscr{L}(p; X_1, ..., X_n)$ and compute the maximum likelihood estimate $\hat{p}$.

(b) Suppose we had data points $x_1, x_2, \ldots, x_n$ where $x_i \in \mathbb{N}$, which were drawn from a Borel distribution. A Borel distribution is discrete, takes parameter μ, and has a PMF (probability mass function, a discrete version of a PDF) of $P(X) = \frac{e^{-\mu X}(\mu X)^{X-1}}{X!}$. Given these data points, find the most likely μ using MLE.

(c) Suppose we had data points $x_1, x_2, \ldots, x_n$ where $x_i \in \mathbb{N}$, which were drawn from a Poisson distribution. A Poisson distribution is discrete, takes parameter λ, and has a PMF of $P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ where $\lambda > 0$ and x is a non-negative integer. Suppose we have an exponential distribution as a prior for λ, with parameter α. Thus $P(\lambda) = \alpha e^{-\alpha\lambda}$. Compute the MLE and MAP for λ. What happens as $n \to \infty$?

Hint: you may assume that the negative of the MLE and MAP functions are convex and thus you can take the derivative to find minima.

3 Estimating x^2 – Bias and Variance

Professor Sahai is trying to estimate some function $f(x), x \in \mathbb{R}$ from noisy points, but has forgotten all machine learning and data regression methods. He has asked you to help him.

Let $f(x) = x^2$. Suppose we are trying to learn $f(x)$, but we're only allowed to make three noisy measurements $(x_1, Y_1), (x_2, Y_2), (x_3, Y_3)$, where for $i \in \{1, 2, 3\}$:

$$x_i = i$$

$$Y_i = f(x_i) + Z_i$$

$$Z_i \overset{\text{i.i.d.}}{\sim} N(0, 1)$$

(a) Suppose we are learning $f(x)$ from $(x_1, Y_1), (x_2, Y_2), (x_3, Y_3)$ using Kernel Ridge Regression. From the following options, what choice of kernel $K : (\mathbb{R}, \mathbb{R}) \to \mathbb{R}$ and regularization parameter λ will minimize our regression model's $bias^2$? Also, find the combinations that maximizes $bias^2$, minimizes variance, and maximizes variance. Explain your answer briefly.

$$K_0(a,b) = 1$$
$$K_1(a,b) = (ab+1)$$
$$K_2(a,b) = (ab+1)^2$$
$$\lambda \in 1, 2, 3$$

(b) Let $\mathbb{P}_0$ be the set of all degree-zero polynomials. In terms of random variables Y_1, Y_2, Y_3, find

$$\hat{p}_0(x) = \underset{p \in \mathbb{P}_0}{\operatorname{argmin}} \sum_{i=1}^{3} (p(x_i) - Y_i)^2$$

(c) Fix $t \in \mathbb{R}$. Suppose we tried to estimate $f(t) = t^2$ using $\hat{p}_0(t)$.
What is the bias of $\hat{p}_0(t)$? What is the variance of $\hat{p}_0(t)$? Express your answers in terms of t.

(d) Let $\mathbb{P}_1$ be the set of all degree-one polynomials. In terms of random variables Y_1, Y_2, Y_3 find

$$\hat{p}_1(x) = \underset{p \in \mathbb{P}_1}{\operatorname{argmin}} \sum_{i=1}^{3} (p(x_i) - Y_i)^2$$

.

Hint 1: Recall that every degree-one polynomial p can be expressed as $p(x) = \vec{w}^T \vec{x}$ where $\vec{x} = \begin{bmatrix} 1 \\ x \end{bmatrix}$ and $w \in \mathbb{R}^2$.

Hint 2: Let $A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$. Then $(A^T A)^{-1} A^T \vec{Y} = \begin{bmatrix} \frac{4}{3}Y_1 + \frac{1}{3}Y_2 - \frac{2}{3}Y_3 \\ -\frac{1}{2}Y_1 + \frac{1}{2}Y_3 \end{bmatrix}$.

(e) Fix $t \in \mathbb{R}$. Suppose we tried to estimate $f(t)$ using $\hat{p}_1(t)$.
What is the bias of $\hat{p}_1(t)$? What is the variance of $\hat{p}_1(t)$? Express your answers in terms of t.

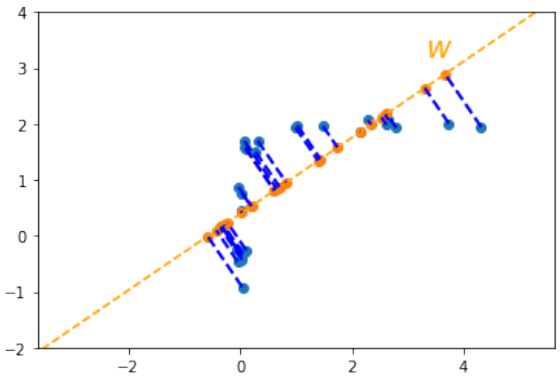
(f) Roughly describe how the bias and variance of $p_0(t)$ and $p_1(t)$ compare as t varies.

4 PCA

Let $X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$ be our data matrix with the datapoints x_i as the rows of the matrix. Assume for this problem that the data is centered, and thus the covariance matrix $\Sigma = \frac{1}{n} X^T X$.

(a) Suppose you were given the unit eigenvectors v_i and eigenvalues λ_i of the covariance matrix. How would you perform PCA to find the best k principal directions and principal coordinates?

(b) Suppose you were given the (compact) SVD of $X = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i u_i v_i^T$. How would you perform PCA to find the best k principal directions and the resulting projected data?



(c) Another way to derive PCA involves finding the direction w that maximizes the variance of the projected data onto w. Specifically, suppose we project x_i onto w, which yields projection coefficient $P_w(x_i)$. We seek to find the w such that we maximize $Var(P_w(x_1), \ldots, P_w(x_n))$.

i. Write the expression for the projection coefficient $P_w(x_i)$.

ii. Find an expression for our objective $Var(P_w(x_1), \ldots, P_w(x_n))$ in terms of X and w. Does this remind you of a certain expression?

iii. What is $\max_w Var(P_w(x_1), \ldots, P_w(x_n))$? How does this relate to PCA?