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?

Solution: Create a diagonal matrix C where C_{ii} is equal to weight c_i. We can see that this problem is equivalent to the sum of the squared components of vector (Xw - y) scaled by the weights c_i. In matrix form, this can be written as $(Xw-y)^T C(Xw-y)$ where y is a vector with components y_i and X is a matrix where x_i is row i of X.

(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?

Solution: Since C is a diagonal matrix with positive values on the diagonal,the matrix has a square root. The square root, $C^{1/2} = \texttt{diag}(\sqrt{c})$. We can therefore write $(Xw-y)^T C(Xw-y)$ as $(Xw-y)^T C^{1/2} C^{1/2}(Xw-y)$ which is equivalent to $\|\hat{X}w-\hat{y}\|_2^2$ where $\hat{X} = C^{1/2}X$ and $\hat{y} = C^{1/2}y$. Further, for the norm of a vector with the vector components x and y: $\| \begin{bmatrix} x \\ y \end{bmatrix} \|_2^2 = \|x\|_2^2 + \|y\|_2^2$
Therefore, we can write: $\|\hat{X}w-\hat{y}\|_2^2 + \lambda\|w\|_2^2$ as $\| \begin{bmatrix} \hat{X}w-\hat{y} \\ \sqrt{\lambda}w \end{bmatrix} \|_2^2$ which is equivalent to:
$\|\hat{X}w-\hat{y}\|_2^2$ where $\hat{X} = \begin{bmatrix} C^{1/2}X \\ \sqrt{\lambda}I_m \end{bmatrix}$ and $\hat{y} = \begin{bmatrix} C^{1/2}y \\ 0 \end{bmatrix}$

(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:

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

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:

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

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

Solution: We first observe that this methods generalizes Ridge Regression. That is, if $v = 0$, we recover Ridge and thus we see that the belief this encodes is that the solution w should be close to some v, just like Ridge should be close to 0.
Now, inspired by this, we can try and reduce this problem to standard Ridge Regression by making the change of variables: $w' = w - v$ to get:

$$\min \|X(w' + v) - y\|_2^2$$
$$\text{s.t. } \|w'\|_2^2 \leq \beta^2$$

Which is equivalent to:

$$\min \|Xw' - (y - Xv)\|_2^2$$
$$\text{s.t. } \|w'\|_2^2 \leq \beta^2$$

Or letting $y' = y - Xv$

$$\min \|Xw' - y'\|_2^2$$
$$\text{s.t. } \|w'\|_2^2 \leq \beta^2$$

Which is percisely Ridge Regression except we have shifted y by a linear combination of our data matrix, effectively recentering our regression. From the homework, we see that this problem has solution:

$$w' = (X^T X + \lambda I)^{-1} X^T y'$$

And then making the appropriate substitutions:

$$w - v = (X^T X + \lambda I)^{-1} X^T (y - Xv)$$

Or

$$w = (X^T X + \lambda I)^{-1} X^T (y - Xv) + v$$

A solution which is in line with the interpretation we had above since we are shifiting the centered solution to the correct region by adding v to it.

(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 $\lambda \neq 0$.

Solution:

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

Substituting back in, we note that expressions inside the inside the parentheses below before inversion have the same non-zero eigenvalues:

$$\begin{aligned}
\Phi^\top(\Phi\Phi^\top + \lambda I)^{-1}y &\stackrel{?}{=} (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top y \\
(\Phi^\top\Phi + \lambda I)\Phi^\top(\Phi\Phi^\top + \lambda I)^{-1}y &\stackrel{?}{=} \Phi^\top y \\
(\Phi^\top\Phi\Phi^\top + \lambda\Phi^\top)(\Phi\Phi^\top + \lambda I)^{-1}y &\stackrel{?}{=} \Phi^\top y \\
\Phi^\top(\Phi\Phi^\top + \lambda I)(\Phi\Phi^\top + \lambda I)^{-1}y &\stackrel{?}{=} \Phi^\top y \\
\Phi^\top y &= \Phi^\top y
\end{aligned}$$

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 $\hat{p}$.

Solution: Let α_H be the number of heads and α_T be the number of tails. Remember that our goal with MLE is to find the model parameter p that maximizes the probability of our samples.

$$\begin{aligned}
\hat{p}_{MLE} &= \arg\max_p \mathscr{L}(p; X_1, \ldots, X_n) \\
&= \arg\max_p P(X_1, \ldots, X_n | p) \\
&= \arg\max_p \prod_{i=1}^{n} \Pr(X_i | p) \\
&= \arg\max_p p^{\alpha_H} (1-p)^{\alpha_T} \\
&= \arg\max_p (\alpha_H \ln p + \alpha_T \ln(1-p))
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial p}(\alpha_H \ln p + \alpha_T \ln(1-p)) &= \frac{\alpha_H}{\hat{p}_{MLE}} - \frac{\alpha_T}{1-\hat{p}_{MLE}} = 0 \\
\hat{p}_{MLE} &= \frac{\alpha_H}{\alpha_H + \alpha_T}
\end{aligned}$$

(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.

Solution: Let $S = \sum x_i$.

$$\begin{aligned}
\hat{\mu}_{MLE} &= \arg\max_\mu \mathscr{L}(\mu; x_1, \ldots, x_n) \\
&= \arg\max_\mu P(x_1, \ldots, x_n | \mu) \\
&= \arg\max_\mu \prod_{i=1}^{n} \frac{e^{-\mu x_i}(\mu x_i)^{x_i - 1}}{x_i!} \\
&= \arg\max_\mu \sum_{i=1}^{n} (-\mu x_i + (x_i - 1)\ln(\mu x_i) - \ln(x_i!)) \\
&= \arg\min_\mu (S\mu - (S-n)\ln(\mu) + \text{constant}) \\
&= \arg\min_\mu (S\mu - (S-n)\ln(\mu))
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \mu}(S\mu - (S-n)\ln(\mu)) &= S + \frac{n-S}{\hat{\mu}_{MLE}} = 0 \\
\hat{\mu}_{MLE} &= \frac{S-n}{S} = 1 - \frac{n}{S} = 1 - \frac{1}{\mu_x}
\end{aligned}$$

where μ_x is the average of our data points.

(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.

Solution: Taking the log likelihood for MLE yields:

$$\begin{aligned}
\ln(P(x_1, x_2, \ldots, x_n)) &= \sum_{i=1}^{n} \ln(P(x_i|\lambda)) \\
&= \sum_{i=1}^{n} -\lambda + x_i \ln(\lambda) - \ln(x_i!) \\
&= -n\lambda + S\ln(\lambda) - \sum_{i=1}^{n} \ln(x_i!)
\end{aligned}$$

where $S = \sum_{i=1}^{n} x_i$. Taking the derivative of the negative of this will get us the MLE.

$$\begin{aligned}
n - \frac{S}{\hat{\lambda}_{MLE}} &= 0 \\
\hat{\lambda}_{MLE} &= \frac{S}{n}
\end{aligned}$$

For MAP we will again calculate log likelihoods:

$$\begin{aligned}
\arg\max_{\lambda} P(\lambda \,|\, x_1, \ldots, x_n) &= \arg\max_{\lambda} \ln(P(x_1, \ldots, x_n \,|\, \lambda) P(\lambda)) \\
&= \arg\max_{\lambda} (-n\lambda + S\ln(\lambda) - \text{constant} + \ln(\alpha) - \alpha\lambda) \\
&= \arg\max_{\lambda} ((-n + \alpha)\lambda + S\ln(\lambda))
\end{aligned}$$

Taking derivatives of the negative of this will then give us the MAP estimate:

$$\begin{aligned}
n + \alpha &= \frac{S}{\hat{\lambda}_{MAP}} \\
\hat{\lambda}_{MAP} &= \frac{S}{n + \alpha}
\end{aligned}$$

Comparing the two forms shows that as $n \to \infty$ the a priori guess becomes negligible, and MAP approaches MLE.

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.

$$K_0(a, b) = 1$$

$$K_1(a, b) = (ab + 1)$$

$$K_2(a, b) = (ab + 1)^2$$

$$\lambda \in 1, 2, 3$$

Solution: K_0, K_1, and K_2 are the degree 0, 1, and 2 polynomial kernels. Since the degree of K_2 matches the degree of the underlying model, we should expect K_2 to learn a model that is closest in expectation to x^2. So among the three kernels, K_2 results in the lowest bias. K_0 results in the greatest bias because it is least expressive.

More expressive kernels have greater variance. So among the three kernels, K_0 results in the lowest variance and K_2 results in the greatest variance.

Regularization attempts to reduce the variance by restricting or simplifying the estimator. This simultaneously increases bias. The largest choice of regularization parameter, $\lambda = 3$, corresponds to the strongest regularization and therefore results in the greatest bias and the lowest variance. $\lambda = 1$ results in the lowest bias and the highest variance.

To minimize $bias^2$, we should pick $\lambda = 1$, and K_2.

To maximize $bias^2$, we should pick $\lambda = 3$ and K_0.

To minimize variance, we should pick $\lambda = 3$, and K_0.

To maximize variance, we should pick $\lambda = 1$ and K_2.

(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

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

Solution:

$$\hat{p}_0(x) = \frac{Y_1 + Y_2 + Y_3}{3}$$

(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.

Solution:

$$\begin{aligned}
\text{Bias} &= \mathbb{E}[\hat{p}_0(t) - f(t)] \\
&= \mathbb{E}[\frac{Y_1+Y_2+Y_3}{3} - t^2] \\
&= \frac{1}{3}\mathbb{E}[Y_1+Y_2+Y_3] - t^2 \\
&= \frac{1}{3}(1^2+2^2+3^2) - t^2
\end{aligned}$$

$$\begin{aligned}
\text{Variance} &= \text{Var}[\hat{p}_0(t)] \\
&= \text{Var}[\frac{Y_1+Y_2+Y_3}{3}] \\
&= \frac{1}{9}(\text{Var}[Y_1]+\text{Var}[Y_2]+\text{Var}[Y_3]) + 0 \\
&= \frac{1}{9}(1+1+1) = \frac{1}{3}
\end{aligned}$$

(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

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

.

<u>Hint 1</u>: 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 $w \in \mathbb{R}^2$.

<u>Hint 2</u>: 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 $(A^TA)^{-1}A^T\vec{Y} = \begin{bmatrix}\frac{4}{3}Y_1 + \frac{1}{3}Y_2 - \frac{2}{3}Y_3\\ -\frac{1}{2}Y_1 + \frac{1}{2}Y_3\end{bmatrix}$.

Solution:

$$\vec{w}^* = (A^TA)^{-1}A^T\vec{Y} = \begin{bmatrix}\frac{4}{3}Y_1 + \frac{1}{3}Y_2 - \frac{2}{3}Y_3\\ -\frac{1}{2}Y_1 + \frac{1}{2}Y_3\end{bmatrix}$$

$$\hat{p}_1(t) = (\vec{w}^*)^T\begin{bmatrix}1\\t\end{bmatrix} = (\frac{4}{3} - \frac{1}{2}t)Y_1 + \frac{1}{3}Y_2 + (-\frac{2}{3} + \frac{1}{2}t)Y_3$$

(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.

Solution:

$$\begin{aligned}
\text{Bias} &= \mathbb{E}\left[\hat{p}_1(t) - f(t)\right] \\
&= \mathbb{E}\left[\hat{p}_1(t)\right] - f(t) \\
&= \mathbb{E}\left[(\frac{4}{3} - \frac{1}{2}t)Y_1 + \frac{1}{3}Y_2 + (-\frac{2}{3} + \frac{1}{2}t)Y_3 - t^2\right] \\
&= (\frac{4}{3} - \frac{1}{2}t)\mathbb{E}[Y_1] + \frac{1}{3}\mathbb{E}[Y_2] + (-\frac{2}{3} + \frac{1}{2}t)\mathbb{E}[Y_3] - t^2 \\
&= (\frac{4}{3} - \frac{1}{2}t)(1^2) + \frac{1}{3}(2^2) + (-\frac{2}{3} + \frac{1}{2}t)(3^2) - t^2
\end{aligned}$$

$$\begin{aligned}
\text{Variance} &= \text{Var}\left[\hat{p}_1(t)\right] \\
&= \text{Var}\left[(\frac{4}{3} - \frac{1}{2}t)Y_1 + \frac{1}{3}Y_2 + (-\frac{2}{3} + \frac{1}{2}t)Y_3\right] \\
&= (\frac{4}{3} - \frac{1}{2}t)^2\text{Var}[Y_1] + (\frac{1}{3})^2\text{Var}[Y_2] + (-\frac{2}{3} + \frac{1}{2}t)^2\text{Var}[Y_3] \\
&= (\frac{4}{3} - \frac{1}{2}t)^2 + (\frac{1}{3})^2 + (-\frac{2}{3} + \frac{1}{2}t)^2 \\
&= \frac{1}{2}t^2 - 2t + \frac{4}{9}
\end{aligned}$$

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

Solution:

Bias: $p_0(t)$ and $p_1(t)$ both have bias with small magnitude when t is near the sampled points. However, if we choose t outside $[1,3]$, we find that both models underestimate $f(t)$, leading to increasingly negative bias.

Variance: $p_0(t)$ has constant variance of $1/3$ for all t. $p_1(t)$ has variance that is minimized at $t = 2$. Since $\text{Var}[p_1(2)] = 4/9$, we find that $\forall t \in \mathbb{R}$:

$$\text{Var}[p_1(t)] > \text{Var}[p_0(t)]$$

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 $\Sigma = \frac{1}{n}X^T X$.

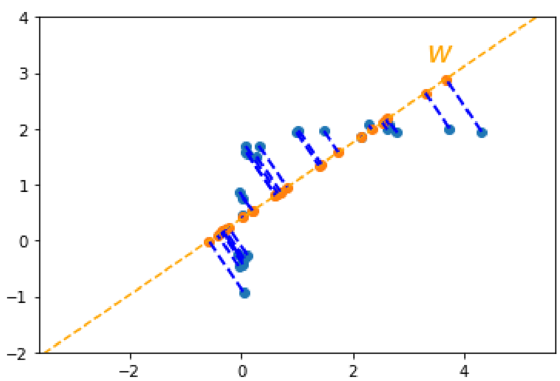
(a) Suppose you were given the unit eigenvectors v_i and eigenvalues λ_i of the covariance matrix. How would you perform PCA to find the best k principal directions and principal coordinates?

Solution: Take the k eigenvectors with the largest eigenvalues. For datapoint x, the projected coefficients for x are simply $x^T v_i$ for all $1 \leq i \leq k$.

(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?

Solution: Note that $X^T X = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^2 V^T$. So, since we're looking for the k eigenvectors corresponding to the largest k eigenvalues of $X^T X$, we simply can take the k right singular vectors v_i corresponding to the k largest singular values σ_i.

For our principal coordinates, note that if $X = U\Sigma V^T$, then $XV = U\Sigma$. Since $XV_{ij} = x_i^T v_j$, our principal coordinates are simply $U\Sigma_k$ where $\Sigma = diag(\sigma_1 \ldots \sigma_k)$.



(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 $P_w(x_i)$.

Solution: The projection of x_i onto w is simply $\frac{x_i^T w}{||w||_2^2} w = \frac{x_i^T w}{||w||_2} \cdot \frac{w}{||w||_2}$. The coefficient is therefore just $\frac{x_i^T w}{||w||_2}$

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?

Solution: Note that the mean of the projected data $\frac{1}{n}\sum_{i=1}^n (P_w(x_i)) = \frac{1}{n}\sum_{i=1}^n \frac{x_i^T w}{||w||_2}$ is 0 since the data is centered.

Thus we have the following expression for this variance:

$$\begin{aligned}
Var(P_w(x_1),\dots,P_w(x_n)) &= \frac{1}{n}\sum_{i=1}^{n}(P_w(x_i) - \frac{1}{n}\sum_{j=1}^{n}(P_w(x_j)))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(\frac{x_i^T w}{||w||_2})^2 \\
&= \frac{1}{n}\frac{||Xw||_2^2}{||w||_2^2} \\
&= \frac{1}{n}\frac{w^T X^T X w}{||w||_2^2}
\end{aligned}$$

This is indeed the rayleigh quotient.

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

Solution:

$$\begin{aligned}
\max_w Var(P_w(x_1),\dots,P_w(x_n)) &= \max_w \frac{w^T X^T X w}{||w||_2^2} \\
&= \max_{||w||_2=1} w^T X^T X w \\
&= \lambda_{max}(X^T X)
\end{aligned}$$

Choosing w to be the eigenvector v_{max} corresponding to $\lambda_{max}(X^T X)$ yields us the maximum. As such, finding the direction w with the maximum variance of the projected data is the same as finding the best principal direction for PCA.