CSE 151A - Homework 4
Ashwin Tyagi. A16228593

---

**Problem 1.** *A simple three-class scenario*

*Solution.* $\mathcal{X} = [-1, 1]$, $\mathcal{Y} = \{1, 2, 3\}$, $\pi_1 = \frac{1}{3}$, $\pi_2 = \frac{1}{6}$, and $\pi_3 = \frac{1}{2}$.

Gaussian Approach: $Y$ for $\max(\pi_Y P_Y(x))$.

$$h^*(\mathcal{X} = [-1, 1]) = \max(\pi_{\mathcal{Y}} P_{\mathcal{Y}}(\mathcal{X}))$$

$$h(-1) = \max((\frac{1}{3})(\frac{7}{8}), (\frac{1}{6})(0), (\frac{1}{2})(\frac{1}{2}))$$

$$h(-1) = \max(\frac{7}{24}, 0, \frac{1}{4}) \longrightarrow \mathcal{Y} = \mathbf{1}$$

$$h(1) = \max((\frac{1}{3})(\frac{1}{8}), (\frac{1}{6})(1), (\frac{1}{2})(\frac{1}{2})$$

$$h(1) = \max(\frac{1}{24}, \frac{1}{6}, \frac{1}{4}) \longrightarrow \mathcal{Y} = \mathbf{2}$$

**Problem 2.** *k classes with a Generative Gaussian Model: linear, spherical, or other quadratic*

*Solution.*

(a) We compute the empirical covariance matrices of each of the k classes, and then set $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$ to the average of these matrices.
**Since all the covariances are set to be equal, the Gaussian Model must be linear**.

(b) The covariance matrices are diagonal, but not identical.
**Although the matricies are diagonal, we do not know if they are positive so it must be some other quadratic.**

(c) There are two classes with covariance matrices that are scalars of the Identity matrix
**Since we are given constant scalars of the Identity matrix, we can say that the Model must be spherical.**

**Problem 3.** *Example of regression with one predictor variable.*

*Solution.*

$$(x, y) = (1, 1), (1, 3), (4, 4), (4, 6)$$

(a) With no knowledge of X:

$$\bar{y} = \frac{1 + 3 + 4 + 6}{4} = 3.5$$

MSE:

$$MSE = \frac{((1 - 3.5)^2 + (3 - 3.5)^2 + (3 - 3.5)^2 + (6 - 3.5)^2)}{4}$$

$$MSE = \frac{((-2.5)^2 + (-0.5)^2 + (0.5)^2 + (2.5)^2)}{4}$$

$$MSE = \frac{12.5 + 0.5}{4} = 3.25$$

(b)

$$MSE = 0.25 \sum_{i=1}^{4} (y^{(i)} - x^{(i)})^2$$

$$MSE = 0.25(0^2 + 2^2 + 0^2 + 2^2) = 2$$

(c) Line of Best Fit

$$y = ax + b$$

$$\bar{y} = 3.5$$

$$\bar{x} = (1 + 1 + 4 + 4)/4 = 2.5$$

$$x = \begin{bmatrix} 1 & 1 \\ 4 & 4 \end{bmatrix} y = \begin{bmatrix} 1 & 3 \\ 4 & 6 \end{bmatrix}$$

$$var(x) = 0.25((-1.5)^2 + (-1.5)^2 + (1.5)^2 + (1.5)^2) = 2.25$$

$$var(y) = 0.25((-2.5)^2 + (-0.5)^2 + (0.5)^2 + (2.5)^2) = 3.25$$

$$cov(x, y) = 0.25((-2.5)(-1.5) + (-1.5)(-0.5) + (1.5)(0.5) + (1.5)(2.5)) = 2.25$$

$$cov(x, y) = \begin{bmatrix} 2.25 & 2.25 \\ 2.25 & 3.25 \end{bmatrix}$$

$$y = \frac{2.25}{2.25} x + (3.25 - (2.25)(1))$$

$$y = x + 1$$

$$MSE = 0.25 \sum_{i=1}^{4} (y^{(i)} - x^{(i)} - 1)^2$$

$$MSE = 0.25((-1)^2 + 1^2 + (-1)^2 + 1) = 1$$

**Problem 4.** *Optimality of the mean*

*Solution.*

(a) Find derivative of $L(s)$

$$L(s) = \frac{1}{n} \sum_{i=1}^{n} (x_i - s)^2$$

$$\frac{d}{ds} L(s) = \frac{-2}{n} \sum_{i=1}^{n} (x_i - s)$$

(b) $L'(s) = 0$

$$0 = \frac{-2}{n}\sum_{i=1}^{n}(x_i - s)$$

$$0 = \sum_{i=1}^{n}(x_i - s)$$

When we minimize the MSE function by setting its derivative to 0, we notice that we are simply summing over all $(x_i - s)$. Therefore, an ideal s value would occur when s is larger than half the values and less than the other half. The best value for this would then have to be the arithmetic mean.

**Problem 5.** *Loss function that corresponds to the total penalty.*

*Solution.* If we predict $\hat{y}$ and the true value is $y$, then the penalty should be the absolute difference, $|y - \hat{y}|$.

$$L(s) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

**Problem 6.** *Writing expressions in matrix-vector form.*

*Solution.*

(a) The average of the $y^{(i)}$ values, that is, $(y^{(1)} + ... + y^{(n)})/n$

$$Out = y^T y/n$$

(b) The n × n matrix whose (i, j) entry is the dot product $x^{(i)} \cdot x^{(j)}$

$$Out = XX^T$$

(c) The average of the $x^{(i)}$ vectors, that is, $(x^{(1)} + ... + x^{(1)})/n$.

$$Out = X^T X/n$$

(d) The empirical covariance matrix.

$$(Xy - (Xy)^T(Xy)/n)(Xy - (Xy)^T(Xy)/n)^T - (X - X^T X/n)(y - y^T y/n)^T$$

**Problem 7.** *Fit (almost) any set of $d + 1$ points*

*Solution.* Let's start with $i = 0$.

$$w \cdot x^{(0)} + b = y^{(0)}$$

$$b = c_0$$

Now we can rewrite every other weight as a difference of $c^{(i)} - c_0$.

$$w \cdot x^{(i)} + c_0 = y^{(i)}$$

$$w \cdot x^{(i)} = c_i - c_0$$

$$w = [0, (c_1 - c_0), (c_2 - c_0), ..., (c_d - c_0)]$$

**Problem 8.** *Ridge Regression*

*Solution.*

(a) What is $L(0)$?
If $\lambda = 0$, then we are not applying a bias to our "perfect" solution. Therefore, $L(0) = 0$ since it too will be "perfect".

(b) As $\lambda$ increases, how does $||w_\lambda||$ behave?
$||w_\lambda||$ will decrease until it is close to 0. This is because the more bias we apply, the further our model will move from the training weights.

(c) As $\lambda$ increases, how does $L(\lambda)$ behave?
$L(0)$ will increase because our loss function will no longer perfectly fit the training data.

(d) As $\lambda$ goes to infinity, what value does $L(\lambda)$ approach?
Since the bias applied by the large lambda effectively removes the $w_\lambda x^{(i)}$ term, we are left with $L(\infty) = \sum_{i=0}^{d}(c_i - (c_0)^2)$.

**Problem 9.** *Discovering Relevant Features in Regression*

*Solution.*

(a) First I am going to parse through the data file so that I have two arrays of size 101 representing my $x$ and $y$ vectors. I will then use the sklearn.linear_model.Lasso package to automatically fit my two vectors. From there it is as simple as grabbing the indicies of the most significant features.

(b) The 10 feature array:

$$[16, 18, 12, 2, 26, 10, 1, 22, 4, 6]$$