HW #3

Uziel Rivera-Lopez

02/20/2023

Question 1

a) w_0, w_1, w_2

$$A = \begin{bmatrix} x_t^2 & x_t & 1|r_t \end{bmatrix}$$

So,

$$\begin{bmatrix} (-2)^2 & 2 & 1|2 \\ 1^2 & 1 & 1|3 \\ 0^2 & 0 & 1|1 \end{bmatrix}$$

Because

$$A \cdot w = \begin{bmatrix} 4 & 2 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} w_2 \\ w_1 \\ w_0 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

So we need to get the inverse of A,

$$A^{-1} = \begin{bmatrix} 4 & 2 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 1/6 & 1/3 & -1/2 \\ -1/6 & 2/3 & -1/2 \\ 0 & 0 & 1 \end{bmatrix}$$

Then, we can solve for w,

$$w = A^{-1} \cdot r = \begin{bmatrix} 1/6 & -1/3 & -1/2 \\ -1/6 & 2/3 & -1/2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} =$$

$$\begin{bmatrix} (1/6 \cdot 2) + (-1/3 \cdot 3) + (-1/2 \cdot 1) \\ (-1/6 \cdot 2) + (2/3 \cdot 3) + (-1/2 \cdot 1) \\ 0 \cdot 2 + 0 \cdot 3 + 1 \cdot 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.8333 \\ -0.1667 \\ 1 \end{bmatrix} = \begin{bmatrix} w_2 \\ w_1 \\ w_0 \end{bmatrix}$$

b) $g(x)$ So using the previous equation we can solve for g(x),

$$g(x) = w_2(x_t)^2 + w_1 x_t + w_0 = 0.8333(x_t)^2 + 0.1667x_t + 1$$

c) R^2
$\Sigma x = -2 + 1 + 0 = -1$
$\Sigma r = 2 + 3 + 1 = 6$
$\Sigma xr = (-2 \cdot 2) + (1 \cdot 3) + (0 \cdot 1) = -4 + 3 = -1$
$\Sigma x^2 = (-2)^2 + 1^2 + 0^2 = 4 + 1 + 0 = 5$
and our n = 3
So we can use the following equation to solve for R^2

$$R^2 = N \cdot \Sigma xr - \Sigma x \cdot \Sigma r \sqrt{N \cdot \Sigma x^2 - (\Sigma x)^2} \cdot \sqrt{N \cdot \Sigma r^2 - (\Sigma r)^2}$$

$$R^2 = 3 \cdot -1 - (-1) \cdot 6 \sqrt{3 \cdot 5 - (-1)^2} \cdot \sqrt{3 \cdot 14 - 6^2}$$

$$R^2 = 0.1071$$

Question 2

a) I would say AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) is a good method of finding good-fit and generalizable model, since it is great with small sample sizes. They are also good at penalizing complex models, so it's a good way to avoid overfitting. Another one would be SRM (Structural Risk Minimization) which is a good way to find a good model, because it sorts the models according to the compexity, as well as other parameters that are used to find the best model.

b) So as we know, bias is a measure of error in our function. While the variance measure the fluctuation around the expected value, EX (This is what we use in stats for expected value). But we cannot have the best of both worlds, like take for example, the variance is 0 so we are hitting our expected value, but we have a very high bias, meaning it's incredlibly wrong. But if you bias is less than say 10% then you are doing pretty good, but your variance will fluctuate by a lot. Bascially it's a inverse relationship, so if you have a high bias, then you will have a low variance, and vice versa.

c) I kinda of explained this in the previous question, but if our bias is high, meaning our model/function is wrong about the samples of data, then our variance will be low, meaning that our data will be close to the expected value. But if our variance is high, then our bias will be low, meaning that our data will be far from the expected value.

Question 3

a) First we need to multiply the z-normalizations, X = 1.5 and Y = 2.4.
$X = [5 * 1.5, 3 * 1.5, 2 * 1.5, 6 * 1.5] = [7.5, 4.5, 3, 9]$
$Y = [7 * 2.4, 5 * 2.4, 7 * 2.4, 5 * 2.4] = [16.8, 12, 16.8, 12]$
Then we can find the mean of X and Y

$\overline{X} = \frac{7.5+4.5+3+9}{4} = 6$
$\overline{Y} = \frac{16.8+12+16.8+12}{4} = 14.4$
Then we can find the covariance of X and Y
$Cov(X,Y) = \frac{\Sigma_{i=1}^{n}(x_i-\overline{X})(y_i-\overline{Y})}{n}$
$Cov(X,Y) = \frac{(1.5)(2.4)+(-1.5)(-2.4)+(-3)(2.4)+(3)(-2.4)}{4}$
$Cov(X,Y) = \frac{3.6+3.6-7.2-7.2}{4}$
$Cov(X,Y) = -1.8$
$\Sigma(X-\overline{X})^2 = \frac{(7.5-6)^2+(4.5-6)^2+(3-6)^2+(9-6)^2}{4} = 5.625$
$\Sigma(Y-\overline{Y})^2 = \frac{(16.8-14.4)^2+(12-14.4)^2+(16.8-14.4)^2+(12-14.4)^2}{4} = 5.76$
So the covariance matrix is:

$$\Sigma = \begin{bmatrix} 5.625 & -1.8 \\ -1.8 & 5.76 \end{bmatrix}$$

b) Find the Joint bivariate distribution of X and Y

$$f(x,y) = \frac{1}{n}\Sigma_{i=1}^{n}\delta(x-x_i, y-y_i)$$

$$f(x,y) = \frac{1}{4}\Sigma_{i=1}^{4}\delta(x-x_i, y-y_i)$$

$$f(x,y) = \frac{1}{4}\delta(x-5,y-7)+\delta(x-3,y-5)+\delta(x-2,y-7)+\delta(x-6,y-5)$$

Question 4

a) Mean imputation involves using data that has missing values, and, since we want to use the data, we decide that the best way to keep the data is by replacing the missing values with the mean of the data. This way we matain our sample size and we can use the data to make predictions. But the problem with this is that it can cause bias in our data, and it can also cause the variance to increase.

b) Non-missing data values:
x_1: [2, 5, 3, 7, 5, 9, 10, 4, 13]
x_2: [6, 4, 3, 1, 8, 12, 23, 3, 5]
x_3: [IN, IN, NY, CA, NY, IL, CA, TX]
Data with missing values:
Finding the mean:

$$\overline{x_1} = \frac{2+5+3+7+5+9+10+4+13}{9} = 6.4444$$

$$\overline{x_2} = \frac{6+4+3+1+8+12+23+3+5}{9} = 7.2222$$

Since we have strings for x_3 we cannot find the mean. So we will need to

x_1	x_2	x_3
2	6	IN
5		
	4	IN
3	3	NY
7		
5	1	CA
	8	NY
9	12	IL
10		
	23	NY
4	3	CA
	5	
13		TX

x_1	x_2	x_3
2	6	IN
5	7.222	
6.444	4	IN
3	3	NY
7	7.222	
5	1	CA
6.444	8	NY
9	12	IL
10	7.222	
6.444	23	NY
4	3	CA
6.444	5	
13	7.222	TX

replace
Now we can replace the missing values with the mean.
We can now take the mean of each x_1 and x_2 and find the missing each missing state, by taking the mean of each current state.

(a) IN: $\frac{2+6+6.444+4}{4} = 4.611$
(b) NY: $\frac{3+3+6.444+8+6.444+23}{4} = 12.472$
(c) CA: $\frac{5+1+4+3}{4} = 3.25$
(d) IL: $\frac{9+12}{2} = 10.5$
(e) TX: $\frac{13+7.222}{2} = 10.111$

For the missing states, we would replace it with the mode of state, which is NY.

x_1	x_2	x_3
2	6	IN
5	7.222	NY
6.444	4	IN
3	3	NY
7	7.222	NY
5	1	CA
6.444	8	NY
9	12	IL
10	7.222	NY
6.444	23	NY
4	3	CA
6.444	5	NY
13	7.222	TX

Question 5

a) Euclidean distance is the length/distance between two points in the Euclidean space, usually in a low dimension as it's not greater in higher. It's often called the shortest path. The Mahalanobis distance is measuring a point with the distribution of many points in a multivariate dimension, usually higher than the Euclidean distance. If we have a dimension say 4 or more, as well as a bunch of points that vary in distance from each other, then we can use the Mahalanobis distance to measure the distance between the points. Because with Euclidean we would need to document the distance between each point, and that would be a lot of work. But if we have a small sample size, then we can use the Euclidean distance to measure the distance between the points.

b) So the shape given the multivariate distribution, $x \sim N(\mu, \Sigma)$, we would expect a circle or an ellipse. This is all due to if the data has a correlation or not which affects the Σ , which is the "variable" making our shape circular or not. The μ is the center of the all the contours that make up this circle or ellipse.

Question 6

a) Using the follow equation:

$$\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

We can compute for class A and class B with vector x = $[1, 2, 0]^T$

$$\mu_A = \begin{bmatrix} 2 & 3 & 1 \end{bmatrix}^T, \Sigma_A = \begin{bmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 2 & 0.2 \\ 0.3 & 0.2 & 1 \end{bmatrix}$$

$$\mu_B = \begin{bmatrix} 0 & 1 & 4 \end{bmatrix}^T, \Sigma_B = \begin{bmatrix} 1.5 & 0.1 & 0.4 \\ 0.1 & 1.8 & 0.6 \\ 0.4 & 0.6 & 2 \end{bmatrix}$$

For Class A:

I) $\frac{1}{(2\pi)^{3/2}|\Sigma_A|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_A)^T\Sigma_A^{-1}(x-\mu_A)\right)$

II) $\frac{1}{(2\pi)^{3/2}|\Sigma_A|^{1/2}}^{[1/2,1/2,-1/2]\cdot\Sigma_A^{-1}(x-\mu_A)}$

III) $\frac{1}{15.7496\cdot|\Sigma_A|^{1/2}}^{[-1.52201]}$

IV)

$$\begin{bmatrix} 0.385/15.7496 & -0.054/15.7496 & -0.053/15.7496 \\ -0.054/15.7496 & 0.262/15.7496 & -0.011/15.7496 \\ -0.053/15.7496 & -0.011/15.7496 & 0.367/15.7496 \end{bmatrix}^{[-1.52201]}$$

For Class B:

I) $\frac{1}{(2\pi)^{3/2}|\Sigma_B|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_B)^T\Sigma_B^{-1}(x-\mu_B)\right)$

II) $\frac{1}{(2\pi)^{3/2}|\Sigma_B|^{1/2}}^{[-1/2,-1/2,2]\cdot\Sigma_B^{-1}(x-\mu_B)}$

III) $\frac{1}{15.7496\cdot|\Sigma_B|^{1/2}}^{[-6.6696]}$

IV)

$$\begin{bmatrix} 0.295/15.7496 & -0.001/15.7496 & -0.032/15.7496 \\ -0.001/15.7496 & 0.274/15.7496 & -0.043/15.7496 \\ -0.032/15.7496 & -0.043/15.7496 & 0.265/15.7496 \end{bmatrix}^{[-6.6696]}$$

x belongs to A

b) We would want to fist calculate the Mahalanobis distance for each class, and then we can compare the two distances to see which class the vector x belongs to.
For Class A:
$D_A^2 = (x-\mu_A)^T\Sigma_A^{-1}(x-\mu_A)$
D_A^2 = [-1.52201]
For Class B:
$D_B^2 = (x-\mu_B)^T\Sigma_A^{-1}(x-\mu_B)$
D_B^2 = [-6.6696]
This after wards determine the probability of the vector x belonging to

class A or B, in a shared covariance matrix.

$$\text{P(A)} = \frac{1}{(2\pi)^{d/2}\det(\Sigma_A)^{1/2}} \exp\left(-\frac{D_A^2}{2}\right)$$
$$\text{P(B)} = \frac{1}{(2\pi)^{d/2}\det(\Sigma_A)^{1/2}} \exp\left(-\frac{D_B^2}{2}\right)$$
$$\text{P(A)} = \frac{1}{(2\pi)^{3/2}\det(\Sigma_A)^{1/2}} \exp\left(-\frac{-1.52201}{2}\right)$$
$$\text{P(B)} = \frac{1}{(2\pi)^{3/2}\det(\Sigma_A)^{1/2}} \exp\left(-\frac{-6.6696}{2}\right)$$
$$\text{P(A)} = \frac{1}{15.7496\cdot\det(\Sigma_A)^{1/2}} \exp\left(0.761005\right)$$
$$\text{P(B)} = \frac{1}{15.7496\cdot\det(\Sigma_A)^{1/2}} \exp\left(3.3348\right)$$

So x belongs to class A. Using a shared covariance matrix allows for simpler calculations, and it seems to suit smaller samples better. But with separate covariance matrices we can see an accurate represenation of each indvidual class, especially if the classes are not similar. If we are using a shared covariance matrix, then we are assuming that the classes are similar, and that the covariance matrix is the same for each class as well as dealing with smaller data, the opposite is true for separate covariance matrices.