Assignment 03

Theoretical Questions

Question 1

You are given with the following polynomial regression equation:

$$g\left(x_t \mid w_2, w_1, w_0\right) = w_2\left(x_t\right)^2 + w_1 x_t + w_0$$

Below is the dataset provided to you where x is an independent variable and r is the dependent variable

- $(x_1, r_1) = (-2, 2)$
- $(x_2, r_2) = (1, 3)$
- $(x_3, r_3) = (0, 1)$

Based on the data and equation provided, calculate following using vector-matrix form (Aw = y) :

a) w_0, w_1, w_2

By plugging in the values of the systems of equations we can calculate the matrix form Y and D

$$y = \begin{pmatrix} \Sigma r_{\mathrm{t}} \\ \Sigma x_{\mathrm{t}} r_{\mathrm{t}} \\ \Sigma {x_{\mathrm{t}}}^2 r_{\mathrm{t}} \end{pmatrix} = \begin{pmatrix} 2+3+1 \\ -2(2)+1(3)+0(1) \\ (-2^2)2+1^2(3)+0^2(1) \end{pmatrix} = \begin{pmatrix} 6 \\ -1 \\ 11 \end{pmatrix}$$

$$D = \begin{pmatrix} 1 & -2 & 4 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Given the matrix form and the inverse given we can set up the equation to be

$$w = (D^T D)^{-1} \dot{y}$$

$$\begin{pmatrix} w0 \\ w1 \\ w2 \end{pmatrix} = \begin{pmatrix} .2 & .05 & .3 \\ .07 & -.03 & .15 \\ .12 & .6 & .3 \end{pmatrix} \begin{pmatrix} 6 \\ -1 \\ 11 \end{pmatrix}$$

Using Matrix Algebra we solve

$$\begin{pmatrix} w0 \\ w1 \\ w2 \end{pmatrix} = \begin{pmatrix} 4.45 \\ 2.1 \\ 3.42 \end{pmatrix}$$

b) Regression Equation $g(x)$
$g(x) = 4.45 + 2.1x + 3.42^2$
c) R-squared $\left(R^2\right)$ value

$$R^2 = 1 - \frac{sse}{sst}$$

$$SSE = \Sigma(y_i - \hat{y})^2 = 267.518$$

$$SST = \Sigma(y_i - \bar{y}) = 2$$

$$R^2 = -132.759$$

Note: You can use the inverse provided below during your calculation

$$\left(D^T \cdot D\right)^{-1} = \begin{bmatrix} 0.2 & 0.05 & 0.3 \\ 0.07 & -0.3 & 0.15 \\ 0.12 & 0.6 & 0.3 \end{bmatrix}$$

Note: Show all steps during your calculations

Question 2

Based on your understanding answer the following questions:

a) Name and explain two methods to select a good-fit and generalized model.
Two models to select a good-fit are cross-validation and regularization.
Cross Validation:

- We plot the total error rate for models with increasing order trained on split data, such that we find the "elbow" in the data set which hopes to find the hidden intersection between the bias and variance unknown to us

Regularization

- By augmenting the error function, regularization involves including lambda * complexity with the normal error to discourage higher order polynomials in the data set. This penalty can be tuned by changing lambda.

b) Explain the Bias-Variance tradeoff.

- The Bias-variance tradeoff demonstrates the strengths and weaknesses of a model with increasing order. When Increasing complexity or order of data we will reduce the bias or error rate with a more flexible model but will also increase the variance as different data set is given. This increases the chances the model will be fit to the noise of the data and not the underlying function. Lower complexity models will be less flexible but be less prone to being fit to the noise

c) Explain what high bias and high variance indicate about your model.

- a model with high complexity and high variance indicates that a model is not only sensitive to the data set, but is not capturing the underlying patterns. This is usually unlikely but may be a result of using the wrong category to predict the target value, such that there is no relationship.

Question 3

Assume that X and Y are bivariately normally distributed variables, you are provided with the z-normalization for both variables: 1.5 for X and 2.4 for Y. Below is the data for both variables:

X	**Y**
5	7
3	5
2	7
6	5

Based on the provided data, calculate the following

a) Covariance Matrix

note: This is calculated using biased in which denominator = N instead of unbiased in which denominator = N-1

mean x = 4

mean y = 6

$$\sigma^2 = \Sigma(x_\mathrm{i} - \bar{x})^2 = 2.5$$

$$\sigma y = 1.58$$

$$\sigma^2 = \Sigma(y_\mathrm{i} - \bar{y})^2 = 4$$

$$\sigma x = 2$$

$$Cov(X, Y) = \Sigma(x - \bar{x})(y - \bar{y})/N = -.5$$

$$CovMatrix(x, y) = \begin{pmatrix} var(X) & cov(x, y) \\ cov(x, y) & var(y) \end{pmatrix} = \begin{pmatrix} 2.5 & -.5 \\ -.5 & 4 \end{pmatrix}$$

$$\rho(x,y) = \frac{cov(x,y)}{\sigma(x)\sigma(y)} = -.158$$

b) Joint bivariate density
given the formula for joint bivariate density

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(z1^2 - 2\rho(z_1z_2) + z2^2)\right)$$

we can plug in the values for our distributions and the z normalization and simplify to

$$f(x,y) = \frac{1}{2\pi(1.58)(2)\sqrt{1-(-0.158^2)}} \exp\left(-\frac{1}{2(1-(-0.158)^2)}(2.25 - 2(-.158)(1.5)(2.4) + 5.76)\right)$$

$$f(x,y) = \frac{1}{20.101} \exp(4.46)$$

Note: Show all steps during your calculations

Question 4

a) Explain Mean imputation

- mean imputation is when missing values in the data are filled with the mean value of that column. This method is not applicable to categorical data such as city names below.

b) Fill out the missing values in following table
mean of x1: 58/9 = 6.44
mean of x2: 65/9= 7.22

x1	**x2**	**x3**
2	6	Indiana
5	7.22	
6.44	4	Indiana
3	3	New York
7	7.22	
5	1	California
6.44	8	New York
9	12	Illinois
10	7.22	
6.44	23	New York
4	3	California
6.44	5	
13	7.22	Texas

Note: Show all steps during your calculations

Question 5

a) Explain the concepts and applications of Mahalanobis distance and Euclidean distance in machine learning. Under what conditions might one be more advantageous than the other?

- Euclidean distance and mahalanobis distance are two ways when finding relationships in a space. Euclidean distance is used to find the distance in between two points in a normal 2d cartesian plane. The mahalanobis is the distance from a point to the center of the 3 dimensional gaussian normal curve. Since this 3d normal distribution accounts for the covariance and deviations of each variable, the mahalanobis distance is weighted by those factors.

b) Given a multivariate distribution, $x \sim N(\mu, \Sigma)$ in 2D what shape do you expect for the contour map and explain how is this shape related to μ and Σ ?

- given different values ofμ and σ the shape of the map changes. As the μ changes the center point of the map moves, when changing the first value it moves the value along the x1 axis and the scond along the x2. When adjusting σ you are able to alter the slope and the size of the curve. Given the 2dimensional view of this curve a positive covariance will slope the bell curve in the same way, this applies for negative covariance as well. The variance of each axis will stretch and shrink the curve so that is wider or narrower.

Question 6

a) Consider a document classification problem where documents are represented as vectors in a high-dimensional space, and the class labels are modeled using a multivariate normal distribution. In this scenario, we have two classes, Class A and Class B.

Mean vector for documents in Class A:

$$\mu_A = [2, 3, -1]^T$$

Covariance matrix for Class A:

$$\Sigma_A = \begin{bmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 2 & 0.2 \\ 0.3 & 0.2 & 1 \end{bmatrix}$$

Mean vector for documents in Class B:

$$\mu_B = [0, 1, 4]^T$$

Covariance matrix for Class B:

$$\Sigma_B = \begin{bmatrix} 1.5 & 0.1 & 0.4 \\ 0.1 & 1.8 & 0.6 \\ 0.4 & 0.6 & 2 \end{bmatrix}$$

a) Calculate the probabilities of the document represented by the vector $x = [1, 2, 0]^T$ belonging to Class A and Class B based on the multivariate normal distribution. Determine the predicted class for the document.

To calculate the probabilities that the vector belongs to class A or Class B we must plug in the vectors into the multivariate equation for each covariance matrix of each classes.

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right))$$

Class A:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma_a|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}([-1,-1,1])^T\Sigma_a^{-1}([-1,-1,1])\right))$$

Simplify

$$p(x|\mu, \Sigma) = \frac{1}{15.74 \cdot 1.26} \exp\left(-1.5249\right)))$$

Continue Simplify

$$p(x|\mu, \Sigma) = 95.14$$

Class B:

$$p(x|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma_b|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}([1,1,-4])^T\Sigma_b^{-1}[1,1,-4])\right))$$

simplify

$$p(x|\mu, \Sigma) = \frac{1}{15.74 \cdot 2.144} \exp\left(-6.668\right)))$$

Continue Simplify

$$p(x|\mu, \Sigma) = 0.36x10^-5$$

b) Now suppose the covariance matrix for both class A and B is Σ_A. Calculate the probabilities of the document represented by the vector $x = [1, 2, 0]^T$ belonging to Class A and Class B. Determine the predicted class for the document. What are the advantages and the assumptions made when using the shared covariance matrix compared to the scenario where separate covariance matrices are used for each class?

When using the same covariance matrix our calculations will become much easier, especially when it comes to inverting matrices which may be computationally expensive. the probability of a would be the same but now lets recalculate for b given the covariance matrix a. Doing this is under the assumption that the variance and covariance of each variable is the same for both classes.

$$p(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma_a|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}([1,1,-4])^T\Sigma_a^{-1}[1,1,-4])\right))$$

$$p(x|\mu,\Sigma) = \frac{1}{15.74 \cdot 1.26} \exp(-10.813)$$

$$p(x|\mu,\Sigma) = 1.0x10^{-}6$$

Note: You may use the following,

$$\Sigma_A^{-1} = \begin{bmatrix} 1.2327 & -0.2767 & -0.3144 \\ -0.2767 & 0.5723 & -0.0314 \\ -0.3144 & -0.0314 & 1.1006 \end{bmatrix}$$

$$\Sigma_B^{-1} = \begin{bmatrix} 0.7043 & 0.0087 & -0.1435 \\ 0.0087 & 0.6174 & -0.1865 \\ -0.1435 & -0.1870 & 0.5847 \end{bmatrix}$$

Practical Questions

Please refer to and answer Questions 7, 8, 9, and 10 in the provided Jupyter Notebook