HW3

LJ Huang

February 25

Question1:

Given the polynomial regression equation $g(x_t|w_2, w_1, w_0) = w_2(x_t)^2 + w_1x_t + w_0$ and the dataset:

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

a) Calculation of w_0, w_1, w_2

Using vector-matrix form $Aw = y$ and the given inverse of $D^T \cdot D$, the weights are calculated as follows:

$$w = (D^T \cdot D)^{-1} \cdot D^T \cdot y = \begin{bmatrix} 0.2 & 0.05 & 0.3 \\ 0.07 & -0.3 & 0.15 \\ 0.12 & 0.6 & 0.3 \end{bmatrix} \cdot \begin{bmatrix} 4 & -2 & 1 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$$

This results in:

$$w = \begin{bmatrix} 3.95 \\ 1.97 \\ 2.52 \end{bmatrix}$$

Thus, $w_2 = 3.95$, $w_1 = 1.97$, and $w_0 = 2.52$.

b) Regression Equation $g(x)$

The regression equation is $g(x) = 3.95x^2 + 1.97x + 2.52$.

c) Calculation of R^2

Given:

$$\begin{aligned}\sum x &= -2 + 1 + 0 = -1\\ \sum r &= 2 + 3 + 1 = 6\\ \sum xr &= (-2 \cdot 2) + (1 \cdot 3) + (0 \cdot 1) = -1\\ \sum x^2 &= (-2)^2 + 1^2 + 0^2 = 5\\ n &= 3\end{aligned}$$

The equation for R^2 is:

$$R^2 = \frac{n \cdot \sum xr - \sum x \cdot \sum r}{\sqrt{(n \cdot \sum x^2 - (\sum x)^2)(n \cdot \sum r^2 - (\sum r)^2)}}$$

Substituting the given values, we find:

$$R^2 = \frac{3 \cdot (-1) - (-1) \cdot 6}{\sqrt{(3 \cdot 5 - (-1)^2) \cdot (3 \cdot 14 - 6^2)}} \approx 0.1071$$

Question2:

a) Two Methods to Select a Good-Fit and Generalizable Model

1. **Cross-Validation:** A technique used to estimate the predictive model's performance in practice, minimizing overfitting and helping to select the model that generalizes well to unseen data.
2. **Regularization Techniques (Lasso and Ridge Regression):** Add penalties on the size of coefficients to reduce overfitting and improve model generalization by either performing variable selection (Lasso) or keeping the model weights small (Ridge).

b) Bias-Variance Tradeoff

The bias-variance tradeoff describes the tradeoff between the error from bias and variance in predictive modeling. High bias can lead to underfitting, missing the complex underlying patterns, while high variance can cause overfitting, capturing noise instead of the underlying data pattern. The goal is to find a balance that minimizes overall error.

c) High Bias and High Variance in Your Model

- **High Bias:** Indicates a too simple model that cannot capture the data's complexity, often leading to underfitting.

- **High Variance:** Indicates a too complex model that captures noise, leading to overfitting and poor performance on unseen data.

Question3:

Given data for variables X and Y, we calculate:

a) Covariance Matrix

The covariance matrix is calculated as follows:

$$\begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix} = \begin{bmatrix} 3.333 & -0.667 \\ -0.667 & 1.333 \end{bmatrix}$$

where $\sigma_X^2 = 3.333$ and $\sigma_Y^2 = 1.333$ are the variances of X and Y, and $\sigma_{XY} = -0.667$ is the covariance between X and Y.

b) Joint Bivariate Density

Given the joint bivariate distribution of X and Y as follows:

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

where δ is the Dirac delta function, and (x_i, y_i) are the given data points. The distribution can be explicitly written as:

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

This representation emphasizes the discrete nature of the given data points in the joint distribution of X and Y.

Question4:

a) Mean Imputation

Mean imputation is a method for handling missing values in a dataset by replacing them with the mean of the available values in the same column for numerical data. It's a straightforward way to maintain the dataset size and structure, particularly useful when the amount of missing data is small. However, this method can potentially reduce the variability of the data and might introduce bias.

b) Filling Out Missing Values

Given the dataset with missing values in columns x_1, x_2, and x_3, we proceed as follows:

For numerical columns x_1 and x_2, we calculate the mean of the available values:

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

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

These means are then used to replace the missing values in x_1 and x_2.

For the categorical column x_3, we cannot compute a mean. Instead, we identify the most frequent category (mode) and use it to replace missing values. If "NY" is the mode, we use it for imputation.

Updated Table:

x1	x2	x3
2	6	IN
5	7.2222	NY
6.4444	4	IN
3	3	NY
7	7.2222	NY
5	1	CA
6.4444	8	NY
9	12	IL
10	7.2222	NY
6.4444	23	NY
4	3	CA
6.4444	5	NY
13	7.2222	TX

Question 5

a) Mahalanobis Distance vs. Euclidean Distance

Euclidean Distance is defined as the square root of the sum of the squared differences between corresponding elements of two vectors. It is used when the dimensions are independent and on the same scale. On the other hand, Mahalanobis Distance considers the correlations between variables and is invariant to scale, useful for detecting outliers and in pattern recognition. Euclidean distance is preferred for physically measured dimensions without scaling issues, while Mahalanobis distance is advantageous for multivariate data, taking into account its covariance structure.

b) Shape of the Contour Map for a Multivariate Distribution

For a multivariate distribution $x \sim N(\mu, \Sigma)$ in 2D, the contour map is elliptical, shaped by the covariance matrix Σ. Diagonal, equal Σ results in circular contours (no correlation, equal variance), while off-diagonal Σ leads to elongated ellipses (indicating correlation and direction of maximum variance). The center of the contours is at the mean μ, depicting the distribution's expected value.

Question 6

a) Separate Covariance Matrices for Class A and B

Given the mean vectors and covariance matrices for Class A and B, and a document vector $x = [1, 2, 0]^T$, we calculate the probability densities based on the multivariate normal distribution. The probability density function for a multivariate normal distribution is given by:

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

Where:

- x is the document vector.
- μ is the mean vector of the class.
- Σ is the covariance matrix of the class.
- Σ^{-1} is the inverse of the covariance matrix.
- $|\Sigma|$ is the determinant of the covariance matrix.
- k is the number of dimensions.

For Class A:

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

For Class B:

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

Calculating the probability densities gives us:

$$p(x|ClassA) \approx 0.01099, \quad p(x|ClassB) \approx 0.000037566$$

Given these densities, the document is more likely to belong to Class A.

b) Shared Covariance Matrix for Class A and B

Assuming Σ_A as the shared covariance matrix for both classes, the probability densities are recalculated. The advantage of using a shared covariance matrix includes simplification of the model and potentially better generalization. The assumption made is that the shape of the distribution is the same across classes, differing only in the mean vectors.

Using the shared covariance matrix, we find:

$$p(x|ClassA, \Sigma_A) = p(x|ClassA), \quad p(x|ClassB, \Sigma_A) \approx 0.000001012$$

Thus, the document is still more likely to belong to Class A, even under the shared covariance assumption.