HW4

LJ Huang

March 17

Theoretical Questions

Question 1

a. Feature Selection vs. Feature Extraction

Feature selection involves choosing a subset of the original features, while feature extraction involves transforming the original features into a new set of features.

Feature Selection:

- **Advantages:** Preserves interpretability, reduces complexity, faster computation.
- **Disadvantages:** May discard relevant information, ignores feature interactions.

Feature Extraction:

- **Advantages:** Captures feature interactions, reduces dimensionality.
- **Disadvantages:** May result in loss of interpretability, computationally intensive.

b. Forward and Backward Selection

Forward Selection: Starts with an empty set of features and iteratively adds features that improve model performance until a stopping criterion is met.

Backward Selection: Starts with all features and iteratively removes features that contribute the least to model performance until a stopping criterion is met.

c. Number of Possible Subsets

Given 20 features and the aim to select 10 features, the number of possible subsets is calculated using the combination formula:

$$\text{Number of Subsets} = \binom{20}{10} = \frac{20!}{10!(20-10)!}$$

Question 2

a. Principal Component Analysis (PCA)

PCA is a dimensionality reduction technique that identifies a set of orthogonal axes (principal components) that capture the maximum variance in the data. It is performed by computing the eigenvectors and eigenvalues of the covariance matrix of the data.

b. Significance of Principal Components

Principal components represent directions of maximum variance in the data. The optimal number of principal components is determined based on the explained variance ratio. Choosing too few components may result in loss of information, while choosing too many components may lead to overfitting.

c. Optimal Choice of Projection Line

The optimal choice of projection line is the line that maximizes the variance of the projected data points. This corresponds to the direction of the first principal component, as it captures the maximum variance in the data.

Question 3

Given the multivariate normal distribution parameters, the distribution of the projected data can be calculated by multiplying the original distribution with the eigen vector matrix.

Question 4

a) Full-Rank Covariance Matrix

A covariance matrix is said to be *full rank* if all its rows and columns are linearly independent, meaning it does not have any redundant or duplicate information. In PCA, a full-rank covariance matrix is important because it ensures that all principal components are unique and capture distinct aspects of variance in the data. This maximizes the effectiveness of dimensionality reduction and avoids redundancy in the components selected.

b) Neglecting Later Eigenvectors

In PCA, it is common to neglect later eigenvectors with smaller eigenvalues because they account for a smaller portion of the variance in the dataset. This practice allows for a focus on the principal components that capture the most significant patterns and structures. However, in cases where small variations are critical to the analysis or when the dataset contains noise that is relevant to the study, considering these less significant eigenvectors may be reasonable.

c) Proportion of Variance Explained

Given the eigenvalues 15, 12, 8, and 3, with a total sum of eigenvalues equal to 50, the proportion of variance explained by the first two principal components is calculated as follows:

$$\text{Proportion of Variance} = \frac{15 + 12}{50} = \frac{27}{50}$$

This suggests that the first two principal components explain a significant portion of the total variance. The choice of how many components to choose depends on the desired level of variance explanation; however, in many cases, selecting components that account for a large proportion of the variance is preferred.

Question 5

Given the covariance matrix $Cov(x)$:

$$Cov(x) = \begin{bmatrix} 0.15 & 1 & 0.2 & 0.08 \\ 0.23 & 0.1 & 0.43 & 0.32 \\ 0.19 & 0.6 & 0.45 & 0.07 \\ 0.3 & 0.4 & 0.5 & 0.07 \end{bmatrix}$$

And the reduced load factors matrix $V_{reduced}$ for the first two dimensions:

$$V_{reduced} = \begin{bmatrix} 0.41 & -0.14 \\ 0.08 & 0.2 \\ 0.03 & -0.07 \\ 0 & 0 \end{bmatrix}$$

We calculate the product of $V_{reduced}$ and its transpose $V_{reduced}^T$, then subtract this from $Cov(x)$ to get the noise matrix Ψ:

$$\Psi = Cov(x) - V_{reduced}V_{reduced}^T$$

After performing the calculations, the noise matrix Ψ is:

$$\Psi = \begin{bmatrix} -0.0377 & 0.9952 & 0.1779 & 0.08 \\ 0.2252 & 0.0536 & 0.4416 & 0.32 \\ 0.1679 & 0.6116 & 0.4442 & 0.07 \\ 0.3 & 0.4 & 0.5 & 0.07 \end{bmatrix}$$

Question 6

a. K-means Clustering

Let's denote the data points as a, b, c, d, e, and the centroids as c_1, c_2, c_3.

Iteration 1:

1. **Calculate Distances:**

	c_1	c_2	c_3
a	9	14	17
b	5	10	31
c	20	25	6
d	2	17	24
e	11	4	37

2. **Assign Clusters:**

$$\text{Clusters: } [1, 1, 3, 1, 2]$$

3. **Update Centroids:**

$$\text{New Centroids: } [25.0, 41.0, 10.0]$$

Iteration 2:

1. **Calculate Distances:**

	c_1	c_2	c_3
a	4	11	16
b	10	5	27
c	15	20	1
d	5	16	21
e	14	3	33

2. **Assign Clusters:**

$$\text{Clusters: } [1, 2, 3, 1, 2]$$

3. **Update Centroids:**

$$\text{New Centroids: } [25.0, 38.0, 10.5]$$

b. Challenges and Initialization Methods

Challenges associated with the dependency on initial choice of centroids include sensitivity to outliers and convergence to local optima. Initialization methods such as k-means++ and hierarchical clustering initialization are used to mitigate these challenges by selecting more representative initial centroids.