Assignment 04

JACK YEUNG B455

Theoretical Questions

Question 1

a) Compare and contrast feature selection and feature extraction techniques for dimensionality reduction, highlighting their respective advantages and disadvantages.

Feature selection and feature extraction are two different techniques to reduce the dimension of a dataset. Feature selection involves picking the most optimal features to represent the data and feature extraction involves forming k new dimension to represent the data in space. Subset selection is the more basic of the approaches which makes it easier to compute, however this process will struggle due to loss of information. This is prevalent in tasks like facial recognition as individual pixels have meaning when combined with others. Feature extraction is more computationally intensive, but can capture information on all dimensions as the k new dimensions are combinations of all d dimensions. It should also be noted that for feature extraction, discrete elements must be prepossessed using one-hot encoding.

b) Describe the forward and backward selection approaches for subset selection. Explain the working principles of each approach and discuss their advantages and disadvantages.

Forward and backward selection use similar heuristics to find the best subsection of k dimensions that represent the results. In forward selection, each variable is evaluated by it's contribution to the model (e.x. reduction in MSE) and then the best is selected. Then after selecting the first best, you continue picking the next optimal dimension to add. This is repeated until the amount of expressively is satisfactory. Backwards selection does the same thing but in reverse, It starts with all dimensions and removes the dimension with the least amount of contributions. Forward selection tends to be much more computationally efficient as you will not have to start with all d dimensions. It should be noted that both of these approaches are greedy and may not find situations where a combination of two factors is greater then each individual contribution combined.

c) Given a dataset with 20 features, if the aim is to reduce the dimensionality to 10 using feature selection, how many possible subsets of features need to be evaluated?

This algorithm is O^2 as you must evaluate all un-selected dimensions. This can be represented as d+(d-1)+(d-2)+...+(d-k). In the case of this we can plug in the values as 20+19+17+..+11 to get the total subsets of 155. Which means we must evaluate 155 subsets before getting to out optimal subset.

Question 2

a) What is PCA and explain how it is performed?

PCA, or principle component analysis is an unsupervised method of feature extraction. The goal of PCA is to project d dimensions onto k principle components with the minimal amount of information lose. This is done by finding the space which maximizes the variation of the data. The process is...

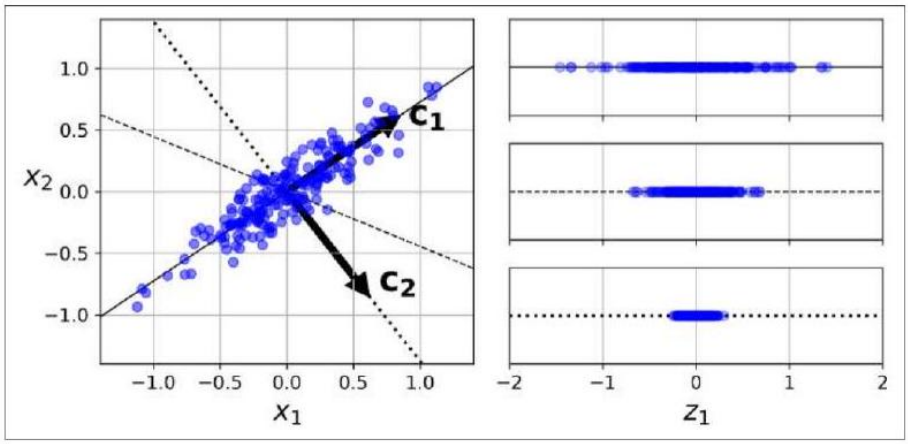
- calculate covariance matrix,
- calculate the eigenvalues and eigenvectors of this matrix.
- Sort by descending order
- Pick the k amount of eigenvectors with the highest eigen-values with subsequent eigenvectors being orthogonal to the selected ones.

b) Discuss the significance of principal components and determine the optimal number to consider. Evaluate the advantages and disadvantages of this selection.

The amount of principle components to select involves considering between how many variables you want to remove and how much variability of the original space you want to represent. This can be done using the scree graph which visualizes the amount of variance is captured when adding each eigenvector. Often times we can select the amount where there is an "elbow" when adding another eigenvector does not contribute as much to variance as previous eigenvectors.

c) In the figure given below a 2D dataset is projected onto three different lines as shown on the right. Which of these lines would be the optimal choice and why?

C1 would be the optimal selection as it expands the data the most, allowing for easier separation between points.



Question 3

Suppose for a dataset X you have a multivariate normal distribution $f(X) \sim N(\mu, \Sigma)$ and is projected onto a new space with the eigen vector matrix W. Calculate the distribution of this
projection given the following values.

$$\mu = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$
$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 2 \end{bmatrix}$$
$$W = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

To project the mean and covariance we have to perform matrix multiplication between the mu and sigma to the W vector.

$$\mu' = W^T \mu$$
$$\begin{pmatrix} 2 \\ 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = 8$$
$$\mu' = 8$$

$$\Sigma' = W^T \Sigma W$$
$$\begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} 0.15 & 1 \\ 0.23 & 0.1 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$
$$\Sigma' = 11$$

In other words when projecting our mulivariate normal distribution onto a singular axis, we have a distribution with mean of 8 and a varaince of 11.

Question 4

a) What does it mean for a covariance matrix to be full rank? In the context of PCA, why is a full-rank covariance matrix important? For a covariance matrix

to be full rank, or to have no zeros. That means that each dimension is linearly independent, this is important in the context of PCA as then there will be all d eigenvectors available.

b) Why is it a common practice to neglect the later eigenvectors with smaller eigenvalues in PCA? In what situations might it be reasonable to consider or not consider these less significant eigenvectors?

The reason later eigenvalues are skipped in PCA is because they provide diminishing returns on the variability expressed in comparison to the full model. In other words the smaller eigenvalues do not separate the data very much making the data more compressed and difficult to differentiate.

c) Given a dataset with a total sum of eigenvalues equal to 50, and the first eigenvalues being $15, 12, 8$, and 3 , determine the proportion of variance explained by the first two principal components. What does this value suggest about the choice of components and how many of these should you choose?

Given the first 4 eigenvalues, we can note that just the first two eigenvectors represent more than half of the variability. It can be noted the rapid increase in subsequent eigenvalues for the next eigenvectors. By selecting these first 4 eigenvalues you will represent 74 percent of the variability of the true data set. However more you want to add is subjective, but each subsequent vector will provide a smaller percentage.

Question 5

Assume you are performing Factor Analysis and you have 4-dimensional data at hand. You are provided with the following data.

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

After performing all the equations, we have received the following load factors for all four dimensions.

$$V = \begin{bmatrix} 0.41 & -0.14 & -0.13 & 0.27 \\ 0.08 & 0.2 & -0.03 & 0.19 \\ 0.03 & -0.07 & -0.21 & 0.11 \\ 0 & 0 & 0 & -0.06 \end{bmatrix}$$

However, you found out that not all dimensions are necessary and chose only the first two dimensions from the given load factors.

Based on the given scenario and all the provided data, calculate the noise matrix (Ψ).

$$\Sigma = VV^T + \Psi$$

Since only 2 dimensions are important we can reduce both the noise and covariance matrix down to 2x2 matrices and insert them into the formula

$$\begin{pmatrix} 0.15 & 1 \\ 0.23 & 0.1 \end{pmatrix} = \begin{pmatrix} 0.41 & -0.14 \\ 0.08 & 0.2 \end{pmatrix} \begin{pmatrix} 0.41 & .08 \\ -.14 & 0.2 \end{pmatrix} + \Psi$$

$$\begin{pmatrix} 0.15 & 1 \\ 0.23 & 0.1 \end{pmatrix} - \begin{pmatrix} 0.1877 & 0.0048 \\ 0.0048 & 0.0464 \end{pmatrix} = \Psi$$

$$\Psi = \begin{pmatrix} -0.0377 & .9952 \\ 0.2252 & 0.0536 \end{pmatrix}$$

Question 6

a) You are tasked with performing K-means clustering on a dataset using three initial random reference vectors (centers) for the clusters, which are given as follows:

- c1 = 30
- c2 = 45
- c3 = 4

You have a dataset with the following data points:

- a = 21
- b = 35
- c = 10
- d = 28
- e = 41

For 2 iterations, answer the following questions:

i) For each iteration, show the distances of each data point from all the reference vectors.

ii) Identify the nearest cluster based on the distance and mention the cluster to which each data point is getting assigned.

Point	c1	c2	c3	closest
A	9	24	17	c1
B	5	10	31	c1
C	20	35	6	c3
D	2	17	24	c1
E	11	4	37	c2

Table 1: Iteration 1

Point	c1	c2	c3	closest
A	7	20	11	c1
B	7	6	25	c2
C	18	31	0	c3
D	0	13	18	c1
E	13	0	31	c2

Table 2: Iteration 2

iii) Showcase the updated values of the reference vectors after assigning data points to the correct clusters after each iteration.
iteration 1:
updated values of clusters after iteration 1 c1 = 28 c2 = 41 c3 = 10
iteration 2:
updated values of clusters after iteration 2 c1 = 24.5 c2 = 38 c3 = 10
b) Explain at least two challenges associated with the dependency on the initial choice of reference vectors in the k-means clustering algorithm. Describe at least two different initialization methods used to mitigate these challenges.

Due to the selection of the initial reference vector, the algorithm is a local search procedure. This means that it can come up with different outputs for different input vectors. There are two main approaches to this

1. take randomly selected k instances as the initial m_i, by running this multiple times with different input, we can use the most common stopping points.

2. Another approach is to take the principal component, divide it's range equally into k different parts, to represent each cluster center.

Practical Questions

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