**Manipal Institute of Technology, Manipal**

**Department of Computer Science and Engineering**

**II semester MTech Computer Science and Engineering**

**Subject: Advanced Machine Learning (CSE5401)**

Time: 10:30 AM -12:30 PM Max. Marks:30

Note:

1 A. Prove that the VC-dimension of rectangle is exactly four. Also prove that no set of five points can be shattered by rectangle. 3M

1 B. Compare feature selection methods with feature extraction methods. 5M

1 C. Given the data points P (3, 2) and Q (4, 1). Evaluate the distance by using any two distance measures. 2M

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | 2 | 3 | 4 | 5 | 6 |
| Y | 3 | 1 | 7 | 8 | 9 |

2 A. Consider the following dataset and predict the regression lines X on Y and Y on X. 4M

2 B. Given the data in Table, discuss the different steps of principal component analysis and compute the covariance matrix and first principal component. 6M

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | 4 | 8 | 13 | 7 |
| Y | 11 | 4 | 5 | 14 |

3 A. We have an agent and a reward, with many hurdles in between to pass. Choose a suitable machine learning technique for the above situation and discuss it with neat diagram.

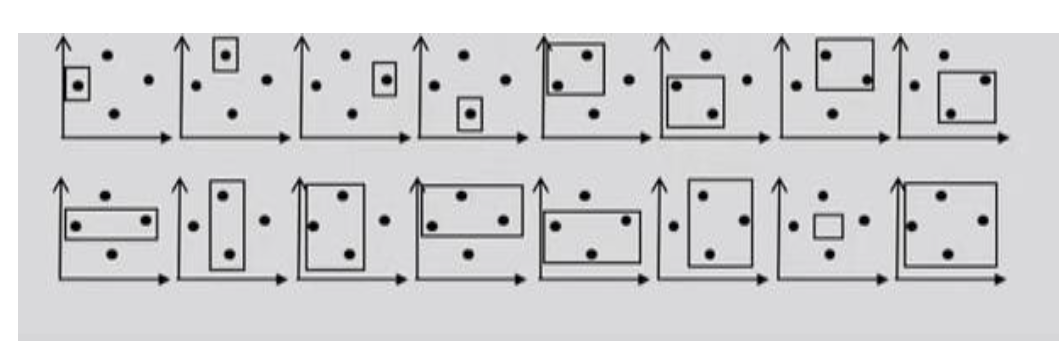
2 M

3 B. Elaborate PCA and LDA methods mathematically. 5 M

3 C. Discuss the advantages and disadvantages of Simple Linear Regression and Multiple Linear Regression models for different types of data analysis tasks. 3 M

**Answers:**

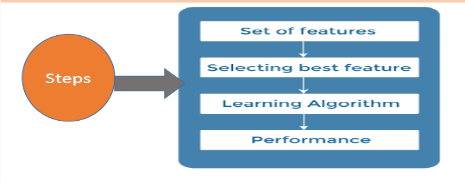
1A. The VC dimension of rectangles is the cardinality of the maximum set of points that can be shattered by a rectangle.16 possible dichotomies. The VC dimension of rectangles is 4 because there exists a set of 4 points that can be shattered by a rectangle. any set of 5 points cannot be shattered by a rectangle.

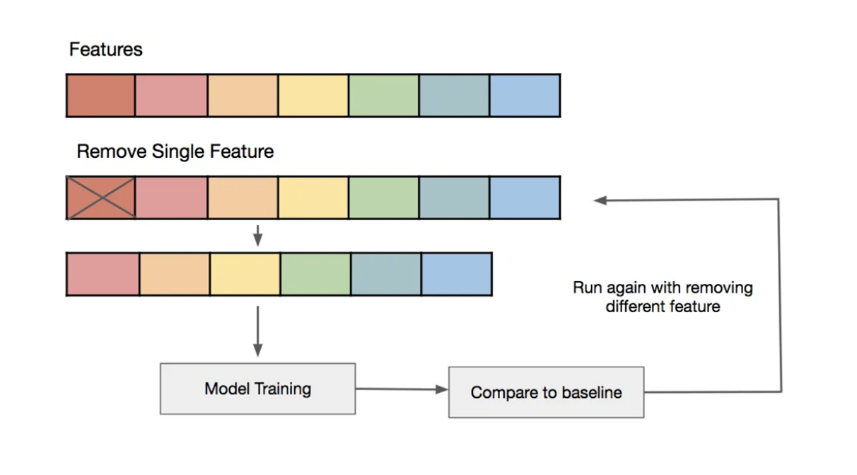
A diagram of a rectangle with arrows and a red dot

Description automatically generated

1B. **feature selection methods**: choosing the important features for the model is known as feature selection. Feature selection is a way of reducing the input variable for the model by using only relevant data to reduce overfitting in the model. Methods are 1. Filter method 2. Wrapper methods. The filter method filters out the irrelevant feature and redundant columns from the dataset by using different metrics. Wrapper methods, also referred to as greedy algorithms train the algorithm by using a subset of features in an iterative manner.

A diagram of a process

Description automatically generated



**Feature extraction:** combining existing features to produce a more useful one

1C. Data points P (3, 2) and Q (4, 1).

**Use Manhattan distance=2 and**

**Euclidean distance= square root of 2.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| X | 2 | 3 | 4 | 5 | 6 |
| Y | 3 | 1 | 7 | 8 | 9 |

2A.

Mean(x)=4,

Mean(y)=5.6,

∑x=20, ∑y=28, ∑xy=131, ∑x2=90, ∑y2=204

**X=1.75552+0.4008Y and Y=-2+1.9X.**

( <https://www.youtube.com/watch?v=RDp6nKRyX1I>)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | 4 | 8 | 13 | 7 |
| Y | 11 | 4 | 5 | 14 |

2B. (Dataset)

**Covariance Matrix= [14, -11**

**-11, 23]**

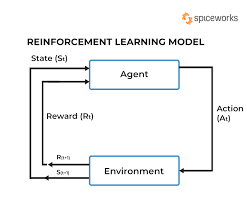
**X2-37x+201=0.**

**first principal component=30.3849 and**

**second=6.6151.**

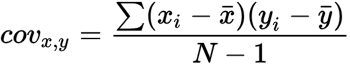
https://www.youtube.com/watch?v=MLaJbA82nzk

3A. **Reinforcement Learning.**



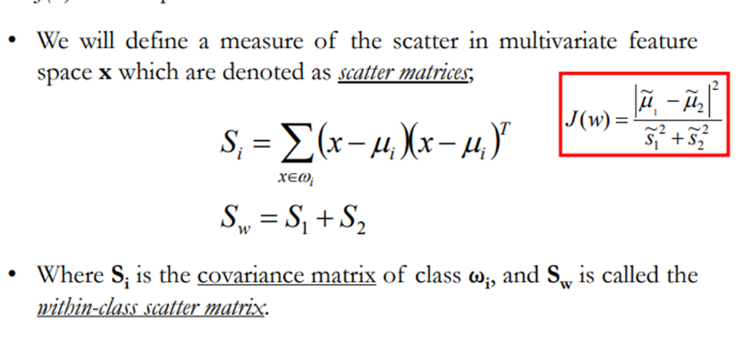
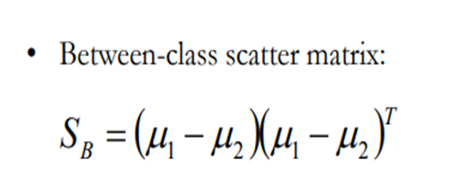
3B. Elaborate PCA and LDA methods mathematically.

**PCA**

* Principal Component Analysis, or PCA, is a dimensionality-reduction method
* Unsupervised method.
* Step 1 - Data normalization.
* Step 2 - Covariance matrix. 
* Step 3 - Eigenvectors and eigenvalues. det(A-λI) = 0, (A-λI)ν = 0
* Step 4 - Selection of principal components.
* Step 5 - Data transformation in new dimensional space.

**LDA**

* LDA is a technique that transforms a set of features or variables into a smaller set of new features, called linear discriminants, that are optimal for separating different classes or categories of the data.
* LDA aims to find the directions or axes that maximize the between-class variance and minimize the within-class variance, and project the data onto those axes



3C. **Linear Regression**

Advantages

1. Simple and easier to interpret.

2. Less complexity

Disadvantages

1. Sensitive to outliers

2. Only captures linear correlation ship between variables.

**Multiple linear Regression**

Advantages

1. It has the ability to determine the relative influence of one or more predictor variables to the criterion value.

2. ability to identify outliers, or anomalies.

Disadvantages

1. It needs high-level mathematics to analyze the data and is required in the statistical program.