UNIVERSITI SAINS MALAYSIA

MACHINE LEARNING CDS503

Topic 9: Dimensionality Reduction

Mohd Halim Mohd Noor, PhD

Outline

- Introduction
- Dimensionality Reduction
- Principal Component Analysis
- Linear Discriminant Analysis
- Filter Methods
- Wrapper Methods

Introduction

- Process of reducing the dimension of the dataset
- A dataset may have hundred of columns (features)
- Reduces the number of columns to a smaller number e.g. 10

Why Dimensionality Reduction?

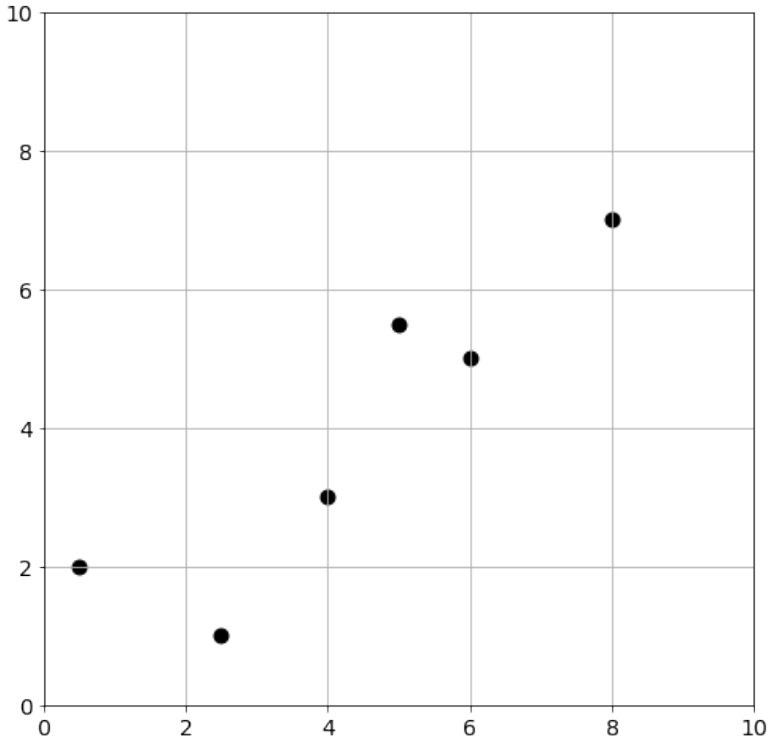
- Remove features that are redundant and not relevant
- Helps in visualizing the data
- Computation time can be reduced thus increasing the speed of our model

Principal Component Analysis

- Linear transformation technique
- Dataset is transformed from its original coordinate to a new coordinate system
- Direction that maximizes the variance in the dataset – bring out strong patterns in the dataset

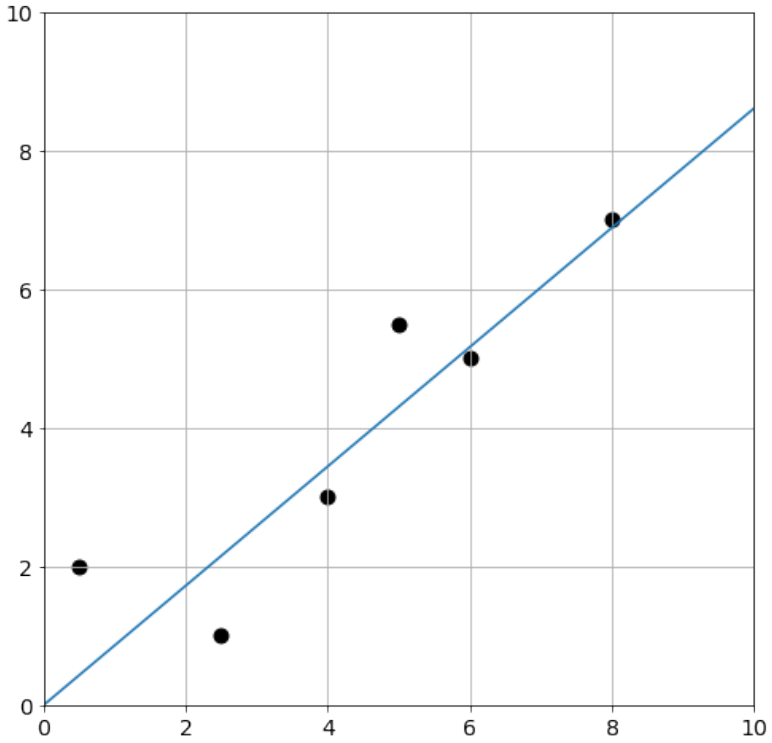
Principal Component Analysis

- Suppose we have the following dataset, which direction maximizes the variance in the data?



Principal Component Analysis

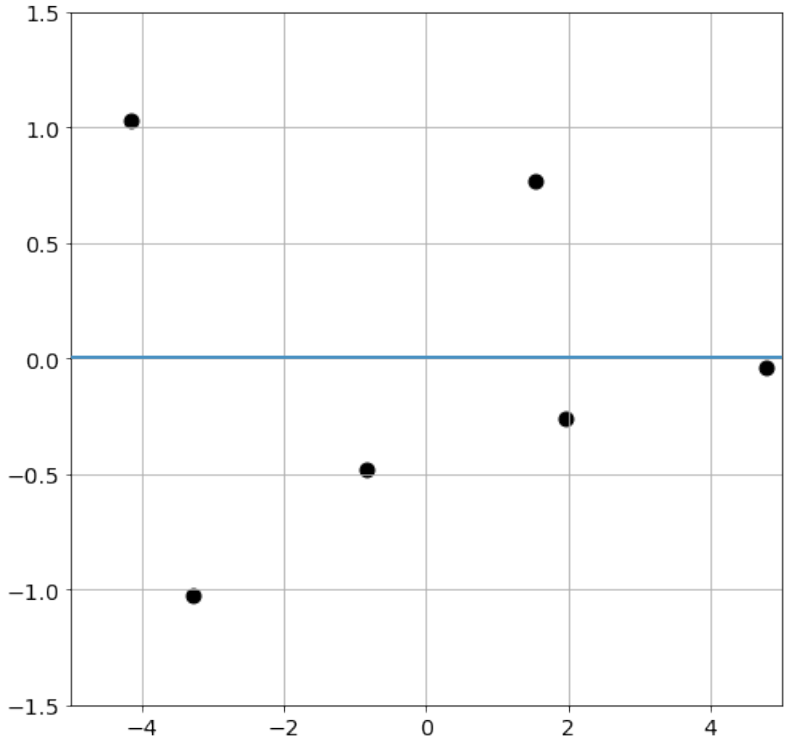
- Suppose we have the following dataset, which direction maximizes the variance in the data?



Blue line cover largest variation in the data

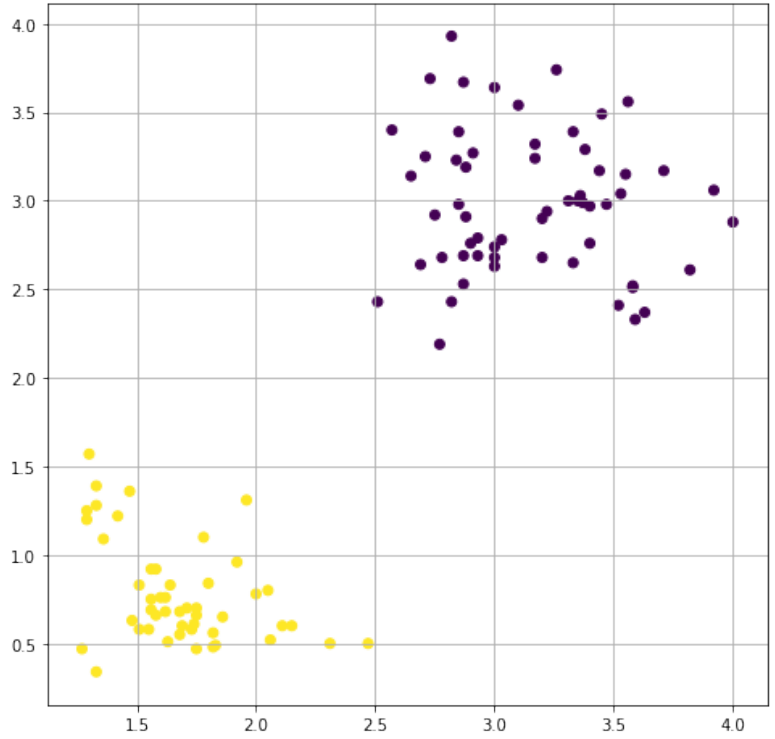
Principal Component Analysis

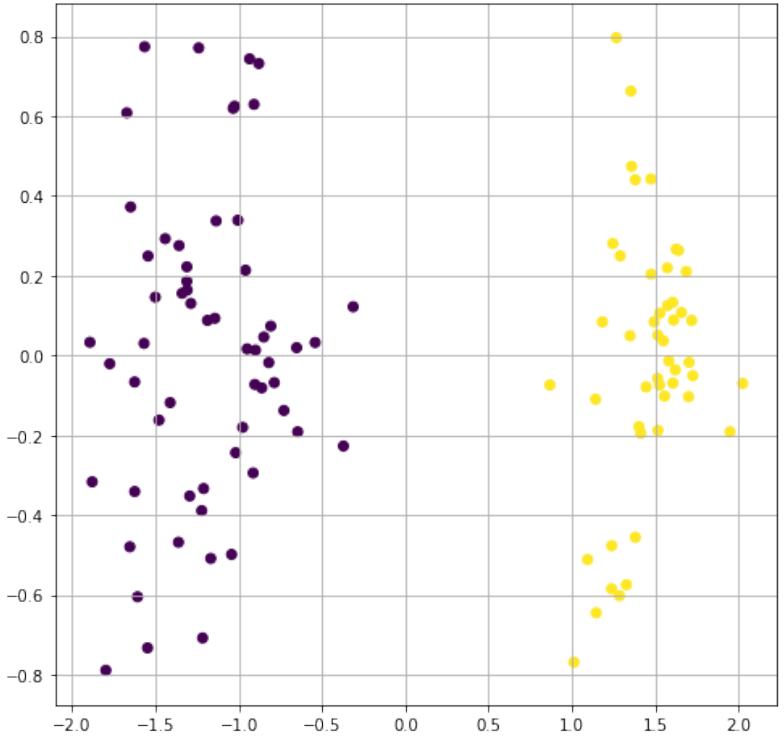
- Suppose we have the following dataset, which direction maximizes the variance in the data?



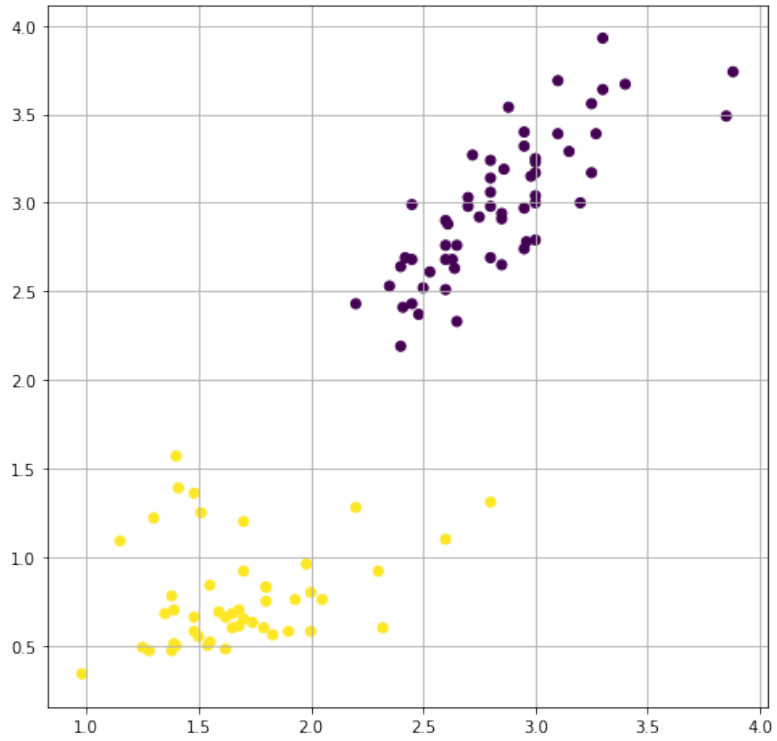
Transform the data such that the data are lined up with blue line

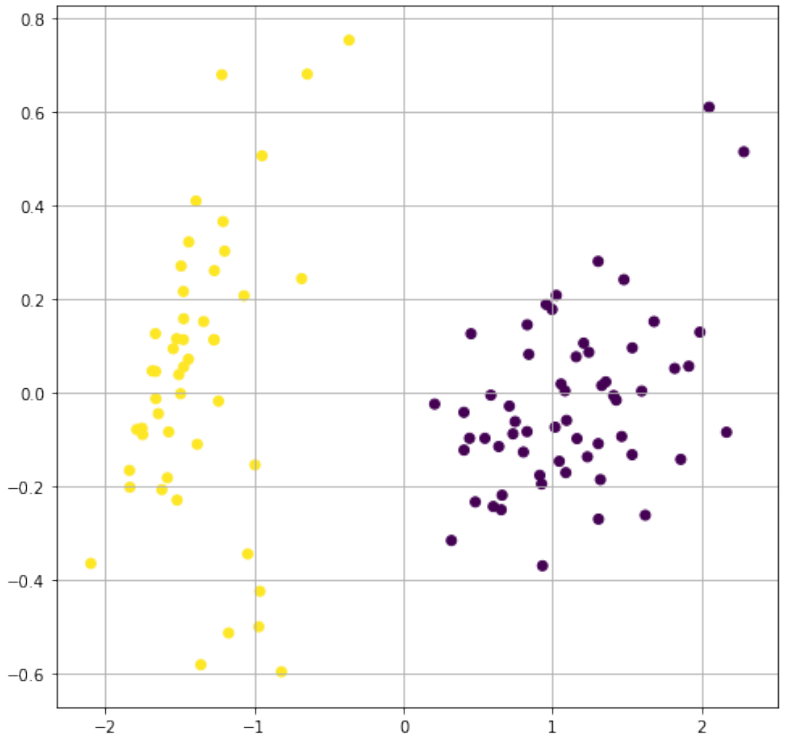
Examples


4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
1.5
2.0
2.5
3.0
3.5
4.0


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

Examples

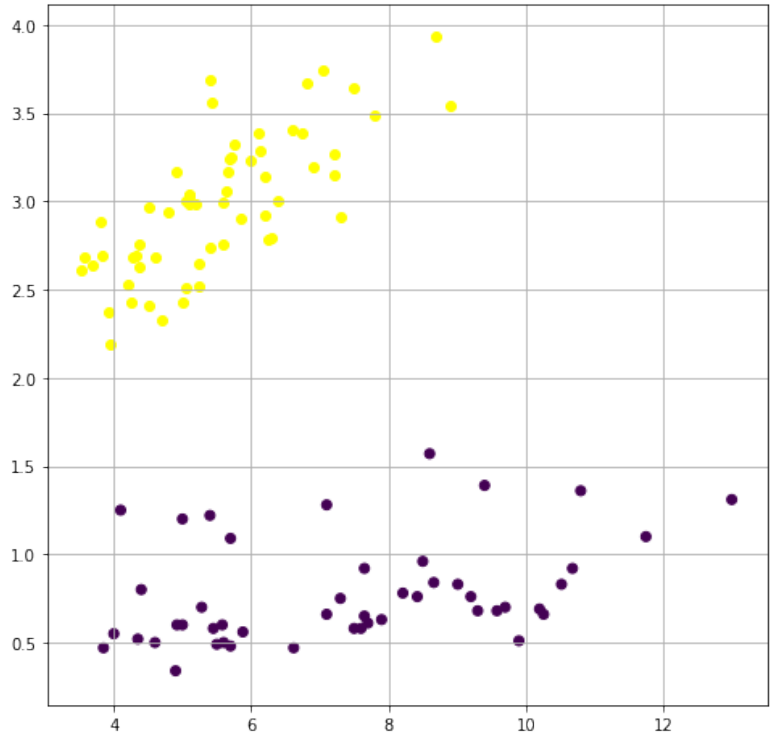

4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−2
−1
0
1
2

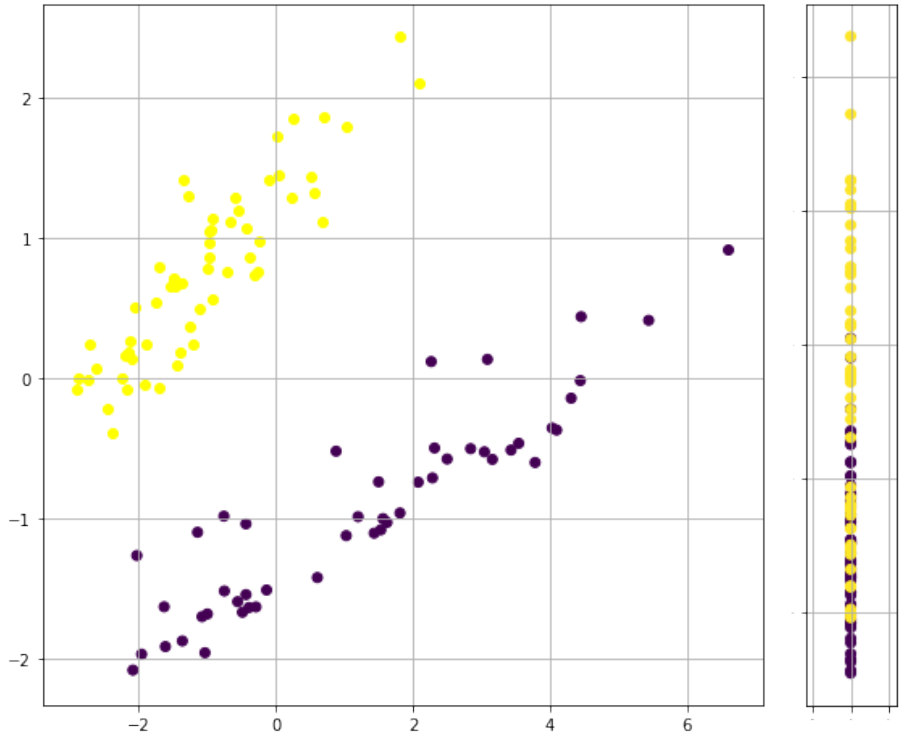
Linear Discriminant Analysis (LDA)

- PCA projects the data in the directions of maximum variance
- Directions of maximum variance may be useless for classification

Linear Discriminant Analysis (LDA)



Before PCA

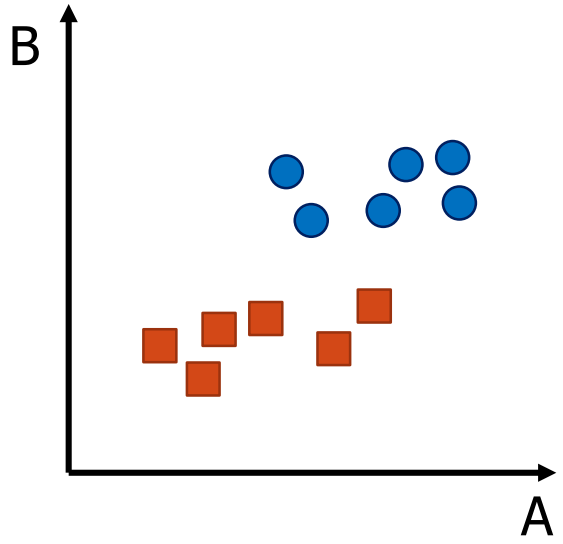


After PCA

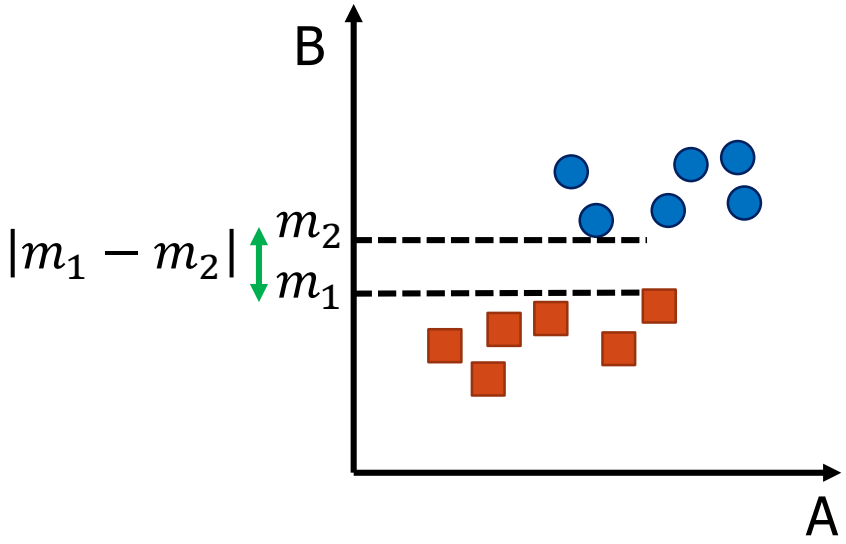
Linear Discriminant Analysis

- Also called Fisher linear discriminant
- Linear transformation technique
- Dataset is transformed to a new coordinate in the direction that maximizes the separation between multiple classes
- How to measure separation between projection of different classes?

Linear Discriminant Analysis

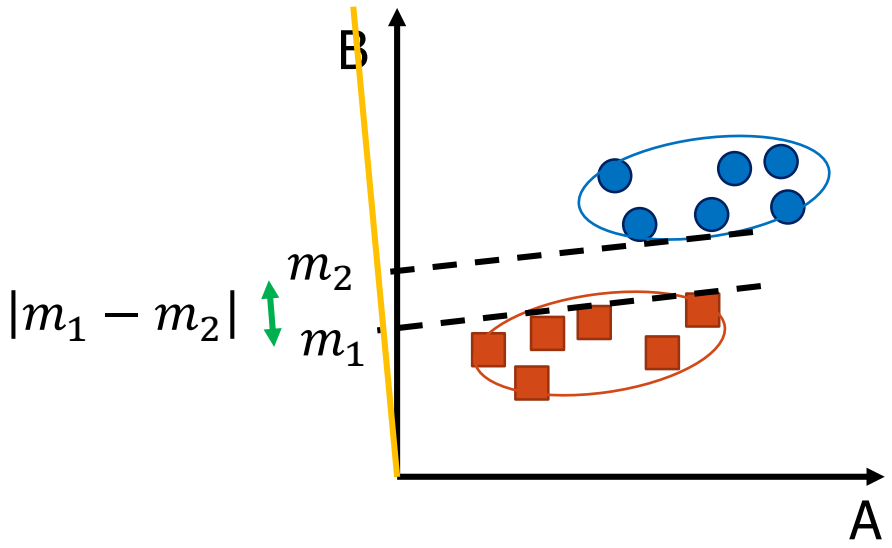


Linear Discriminant Analysis



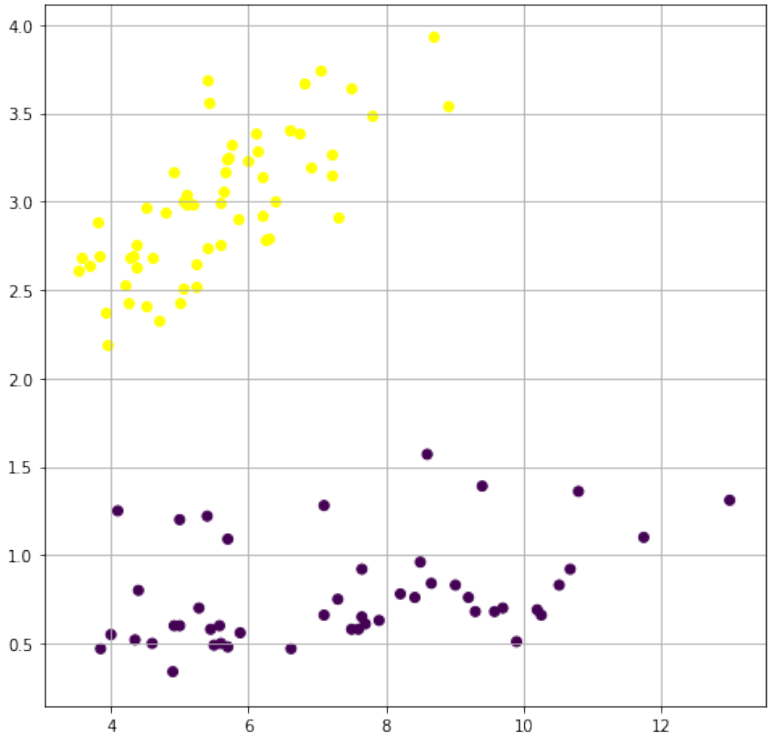
Difference between $|m_1 - m_2|$ seems a good measure of separation

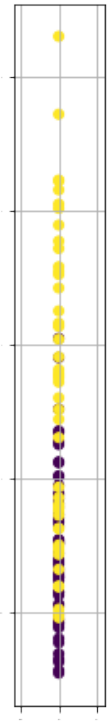
Linear Discriminant Analysis



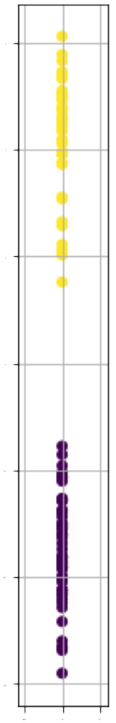
Difference between $|m_1 - m_2|$ seems a good measure of separation

Linear Discriminant Analysis (LDA)





PCA



LDA

Univariate Feature Selection

- Examines each feature individually to determine the strength of the relationship of the feature with the response variable
- Good for gaining a better understanding of the data (but not necessarily for optimizing the feature set for better generalization)

Feature Selection

Age	Attendance	Test 1	Test 2	Project	Final	Grade
28	85	87	83	90	82	A
35	90	79	81	81	78	B
25	90	71	65	75	67	C

- Features = {Age, Attendance, Test_1, Test_2, Project, Final}
- Subset
- {Age, Attendance, Final}
- {Attendance, Test_1, Project}
- {Test_1, Test_2, Project, Final}

Feature Selection

- Filter Methods
  - Use of statistical measure to assign scoring (ranking) to each feature
- Wrapper Methods
  - Use of predictive model as a black box to evaluate the features and assign scores based on predictive model accuracy
  - Predictive mode e.g. Decision Tree, Naïve Bayes etc.

Filter Methods

- Find unique features that contain useful information about the response variable (target) in the data based on statistical methods

- Mutual information
- ANOVA

Filter Methods

- Mutual information
- Measures the dependency between two variables, x and y
- Classification and Regression tasks

Filter Method

- ANOVA
- Measures the difference between two or more groups of data
- Classification and Regression tasks

Wrapper Methods

- Sequential Forward Selection
- Suppose F is a feature set of input dimensions, x_i, $i = 1, \ldots, d$
- Validation error, $E(F)$ when only the features in F are used
- Start with $F = \emptyset$
- At each step, for all possible x_i, train the model on the training set and calculate $E(F \cup x_i)$ on the validation set
- Choose the feature x_j that causes the least error

$$j = \arg\min_i E(F \cup x_i)$$

- add x_j to F if $E\left(F \cup x_j\right) < E(F)$

Wrapper Methods

- Sequential Backward Selection
- Suppose F is a feature set of input dimensions, x_i, $i = 1, \ldots, d$
- Validation error, $E(F)$ when only the features in F are used
- Start with $F = \neg\emptyset$ (all features)
- At each step, for all possible x_i, train the model on the training set and calculate $E(F - x_i)$ on the validation set
- Choose the feature x_j that causes the least error

$$j = \arg\min_i E(F - x_i)$$

- remove x_j from F if $E\left(F - x_j\right) < E(F)$

UNIVERSITI SAINS MALAYSIA

End