

Data pre-processing

Dr Mohamed Bennasar

Outline

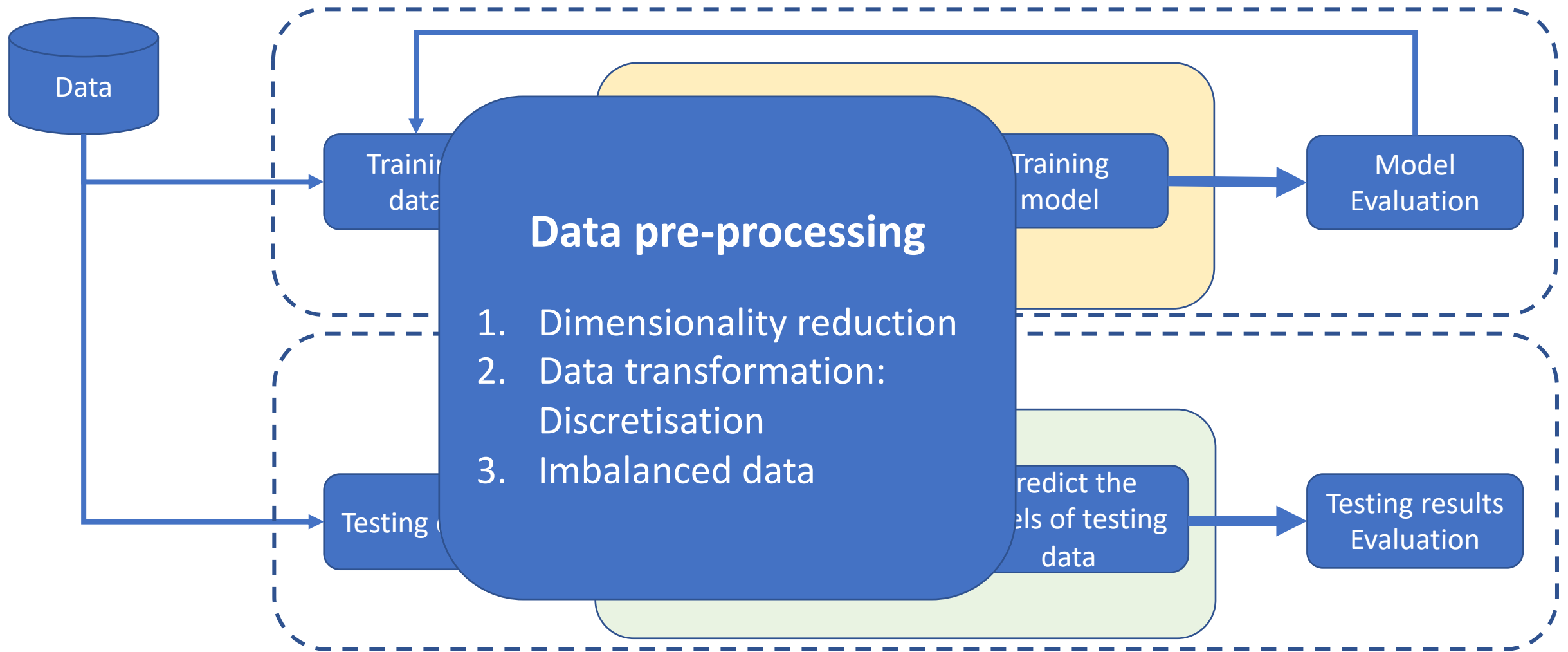
Data Pre-processing

Dimensionality Reduction

Data transformation: Discretisation

Imbalanced Data

Supervised ML pipeline



Why do we need Pre-processing stage?

1. Real-world datasets are highly likely to include noise, missing values, inconsistency, and redundant and irrelevant information.
2. These problems are due to the data collection process involving multiple, heterogeneous sources.
3. Low-quality data will lead to poor machine learning model performance or make the trained model ungeneralisable.
4. Pre-processing is also vital for some machine learning algorithms that need the data to be in a particular format.

Dimensionality reduction

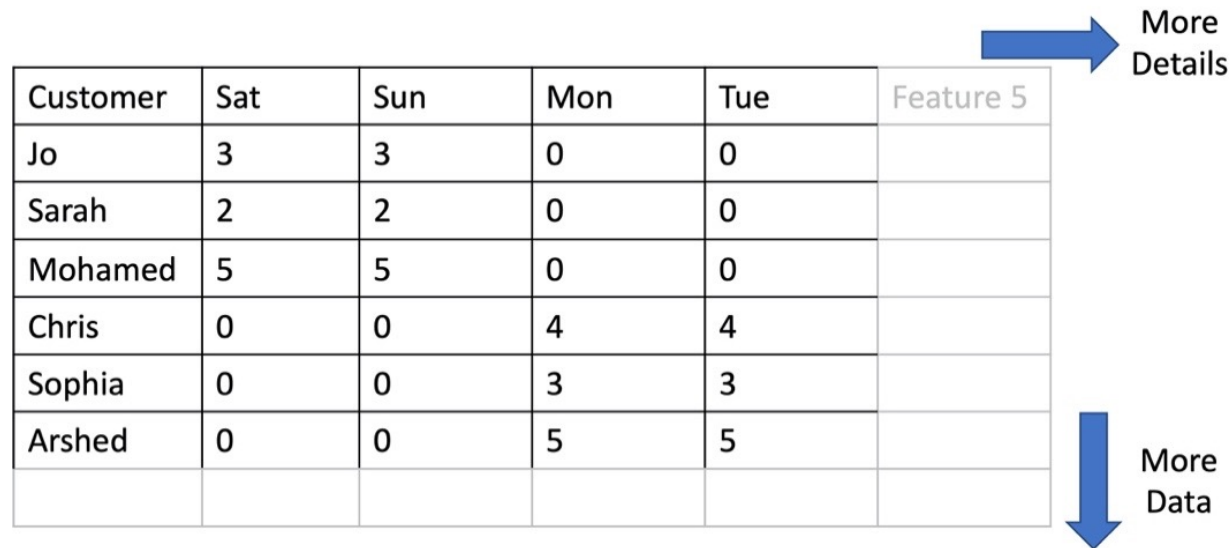
Dimensionality reduction

Dimensionality reduction is about discovering a more compact dimension space to represent the data – in other words, we would like to keep all the rows and reduce the number of columns.

Sex	Age	Weight	Hypertensive	Hypercholesterolemia	Smokeing	Diabetic	Family History	Diagnosis
F	77.00	73.00	1.00	1.00	0.00	0.00	0.00	Neg
F	55.00	70.00	1.00	1.00	0.00	0.00	1.00	Neg
F	67.00	63.00	1.00	1.00	0.00	0.00	0.00	Pos
F	51.00	89.00	1.00	1.00	0.00	1.00	1.00	Neg
F	72.00	56.00	1.00	1.00	0.00	1.00	0.00	Pos
M	67.00	76.00	0.00	1.00	0.00	0.00	0.00	Neg

Dimensionality reduction

Can Increasing the number of features should add more information about each data point (details) and lead to better classification performance?



Customer	Sat	Sun	Mon	Tue	Feature 5
Jo	3	3	0	0	
Sarah	2	2	0	0	
Mohamed	5	5	0	0	
Chris	0	0	4	4	
Sophia	0	0	3	3	
Arshed	0	0	5	5	

High dimensional data

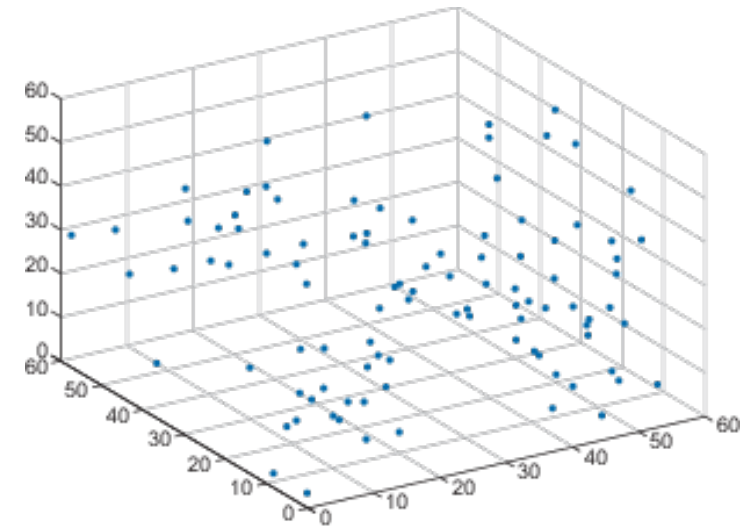
- The inclusion of more features may lead to worse performance if the features are redundant or irrelevant.
- High dimensional data can lead to overfitting.
- Curse of dimensionality

Curse of dimensionality

This problem happens when the number of features is high while the number of observations is low.



Single dimensional data

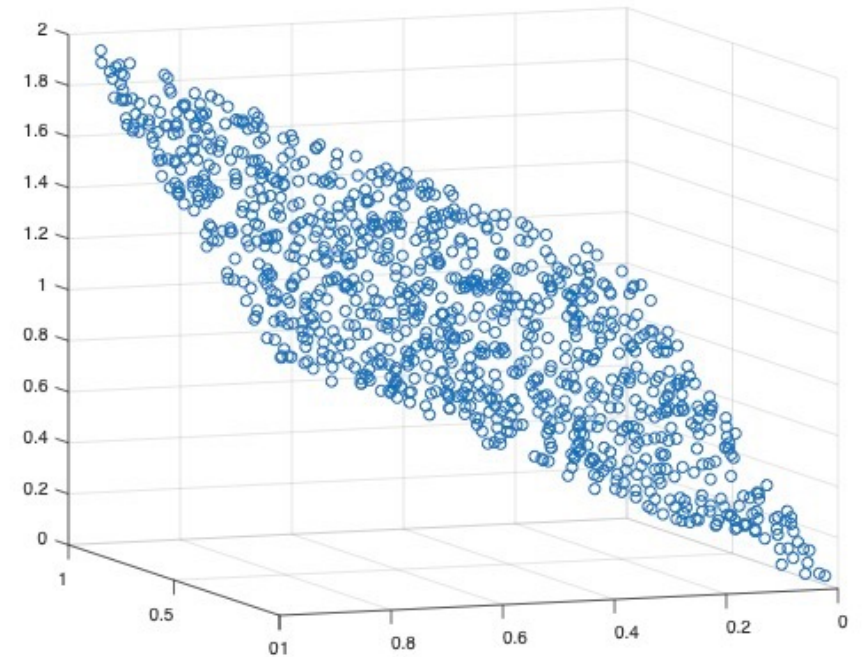


Three dimensional data

Imagine the increase in the distance between the data points if we use 100 features with only 100 data points, which will lead to data fragmentation three features.

Redundant features

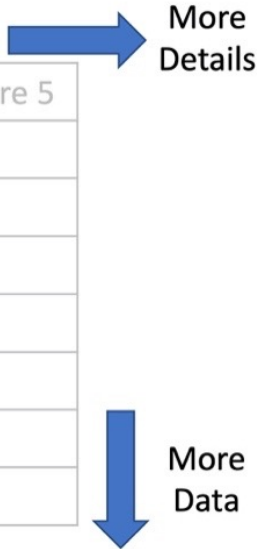
Assumption: We have a 2-D matrix of data. The data lie on 2-D plan which a sub-space of the 3-D space.



Redundant features

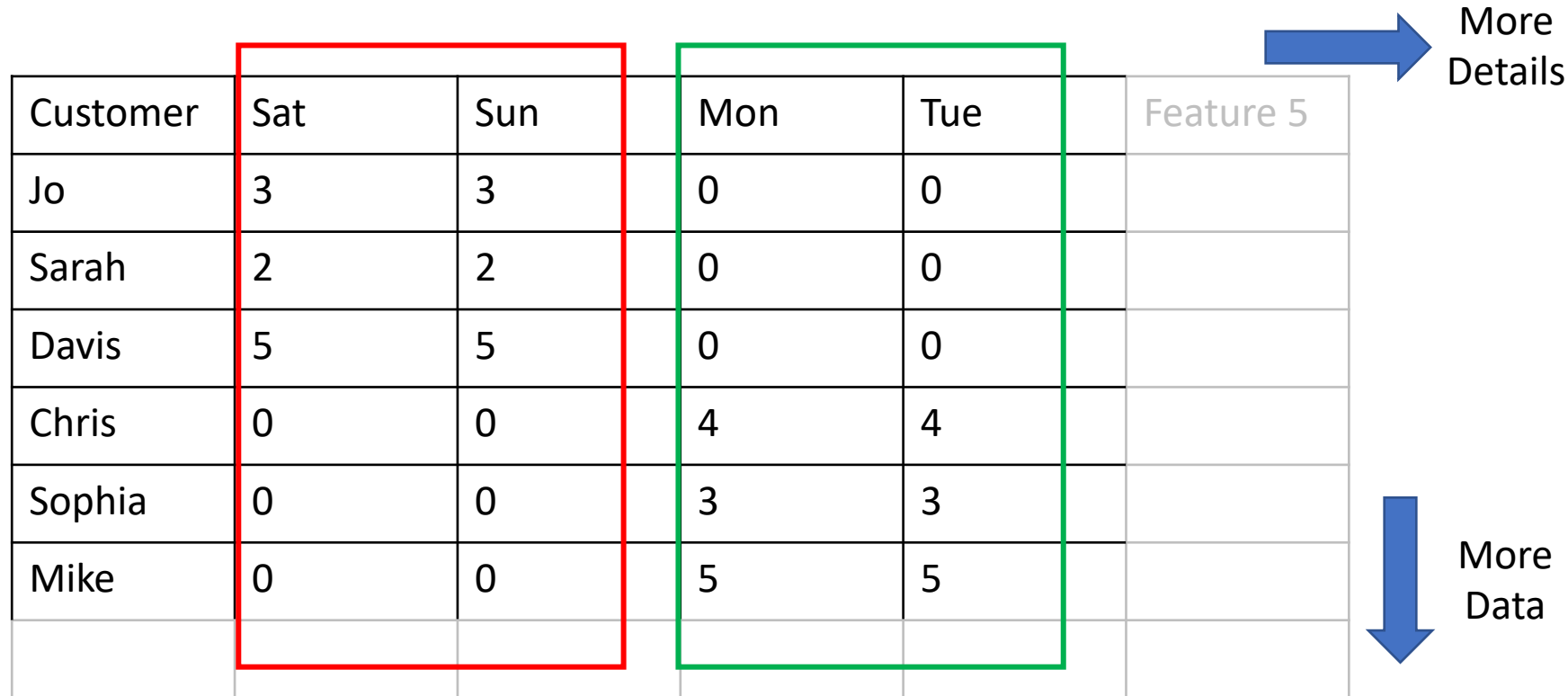
Assume we have the data set shown in the table below.

Customer	Sat	Sun	Mon	Tue	Feature 5
Jo	3	3	0	0	
Sarah	2	2	0	0	
Mohamed	5	5	0	0	
Chris	0	0	4	4	
Sophia	0	0	3	3	
Arshed	0	0	5	5	



Redundant features

Customer	Sat	Sun	Mon	Tue	Feature 5
Jo	3	3	0	0	
Sarah	2	2	0	0	
Davis	5	5	0	0	
Chris	0	0	4	4	
Sophia	0	0	3	3	
Mike	0	0	5	5	



Dimensionality Reduction : Reducing the number of features (columns), while preserving the important information within the data by mapping (projecting) the data into a *lower-dimensional* space.

Why is dimensionality reduction needed ?

- Discover hidden correlation between features.
- Remove redundant and noisy features.
- Have a better interpretation and visualisation.
- Reduce the storage space.
- Reduce the processing time.
- May enhance the performance of the learning algorithms, and reduce the complexity.

Examples of redundant features

- Images
 - Not all the pixels are important.
- Text
 - Not all the words are keywords.
- Time series
 - Not every single reading carry significant information.

Dimensionality techniques

Dimensionality can be reduced by:

- **Feature Extraction** : Generate a combination of the original features using a linear or non-linear transformations.
- **Feature Selection** : Selecting a subset of features which is most useful for the classifier training process from the available features.

Types of Feature selection methods

1. **Filter**: Classifier-independent methods rank features according to their relevance to the class label
2. **Wrapper**: Classifier-dependent methods that search the feature space, and test all possible subsets of feature combinations by using the prediction accuracy of a classifier as a measure.
3. **Embedded**: Classifier-dependent methods that are combined with the learning stage

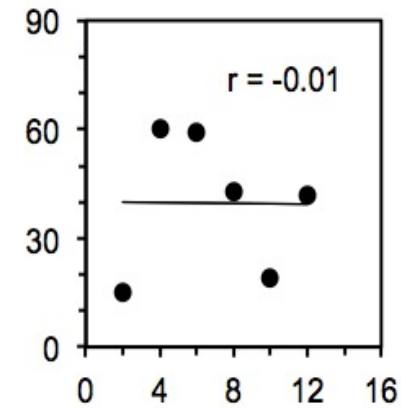
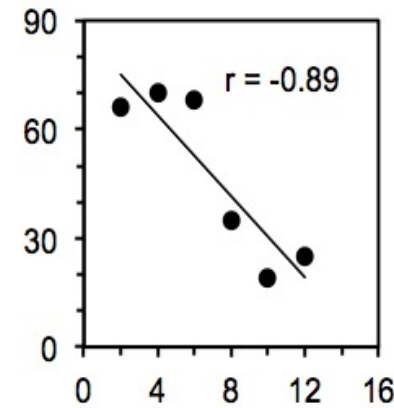
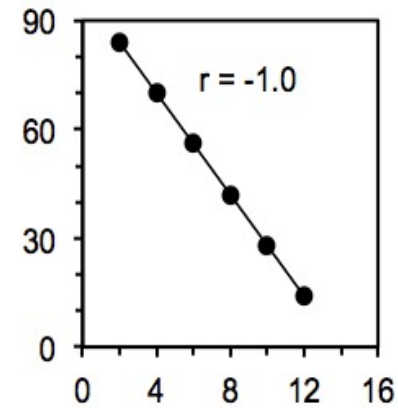
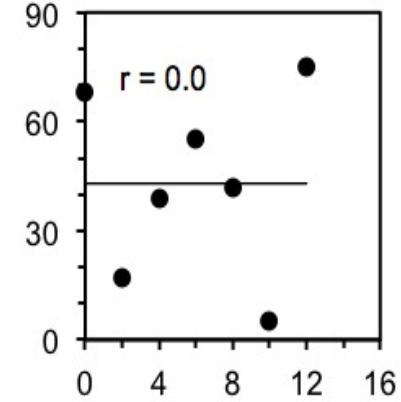
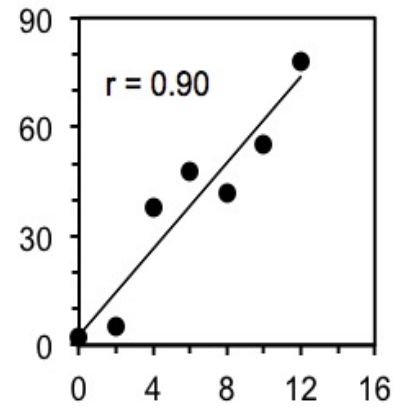
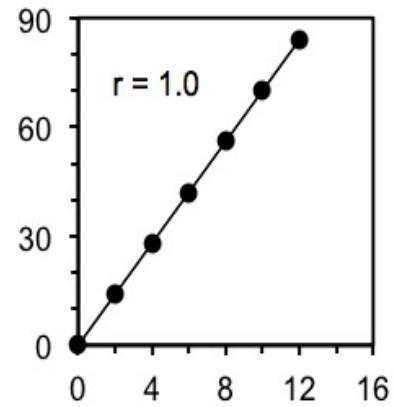
Pearson correlation coefficients

1. Correlation is a technique for measuring the relationship between two quantitative variables.
2. The Pearson correlation coefficient, r , can take a range of values from +1 to -1
3. The value of 0 indicates that there is no association between the two variables.

The Pearson correlation coefficient, r , is calculated using the following equation:

$$r_{x,y} = \frac{\sum_{i=1}^M (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^M \sqrt{(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

Pearson correlation coefficients



Pearson correlation coefficients

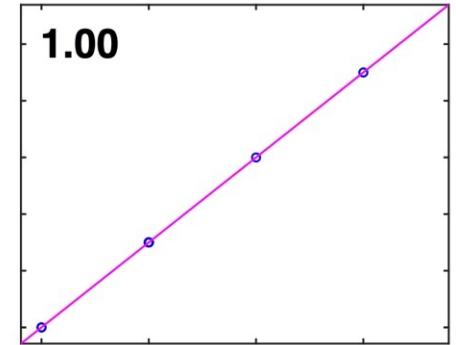
The correlations between each feature and the class label C are

F1	F2	F3	C
11	64	1	2
8	1	1	1
11	64	0	2
14	729	1	3
17	100	0	4
11	64	1	2
8	1	0	1
11	64	1	2

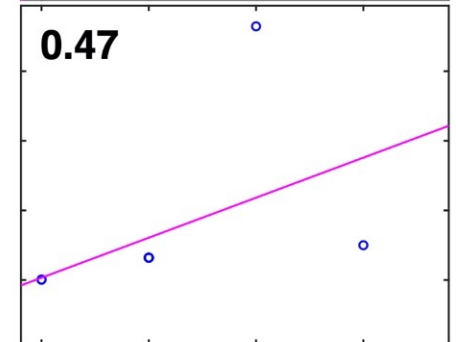
Pearson correlation coefficients

Therefore, F1 should be retained, F3 should be discarded, and Pearson doesn't say much about whether F2 is useful or not.

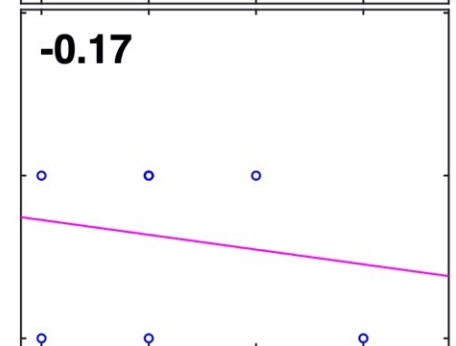
$r_{f1,c}$



$r_{f2,c}$



$r_{f3,c}$



Disadvantages of Pearson correlation coefficients

- It may fail to detect the nonlinear relationship for the second feature.
- It also assumes that the data is normally distributed, which is not always the case.

Normalised Information Gain (IG)

This is a simple method to measure the significance of features based on information theory.

The first concept is Entropy which measures the amount of uncertainty about a variable.

$$H(F) = - \sum_{i=1}^N p(f_i) \log(p(f_i))$$

Mutual information (MI)

Mutual information (MI) between two variables is defined as the amount of information which is shared between the variables.

$$I(F; C) = H(C) + H(F) - H(C, F)$$

To keep its value between 0 and 1, 0 indicates no association, while 1 indicates high association.

$$NI(F; C) = \frac{I(F; C)}{H(C, F)}$$

Example

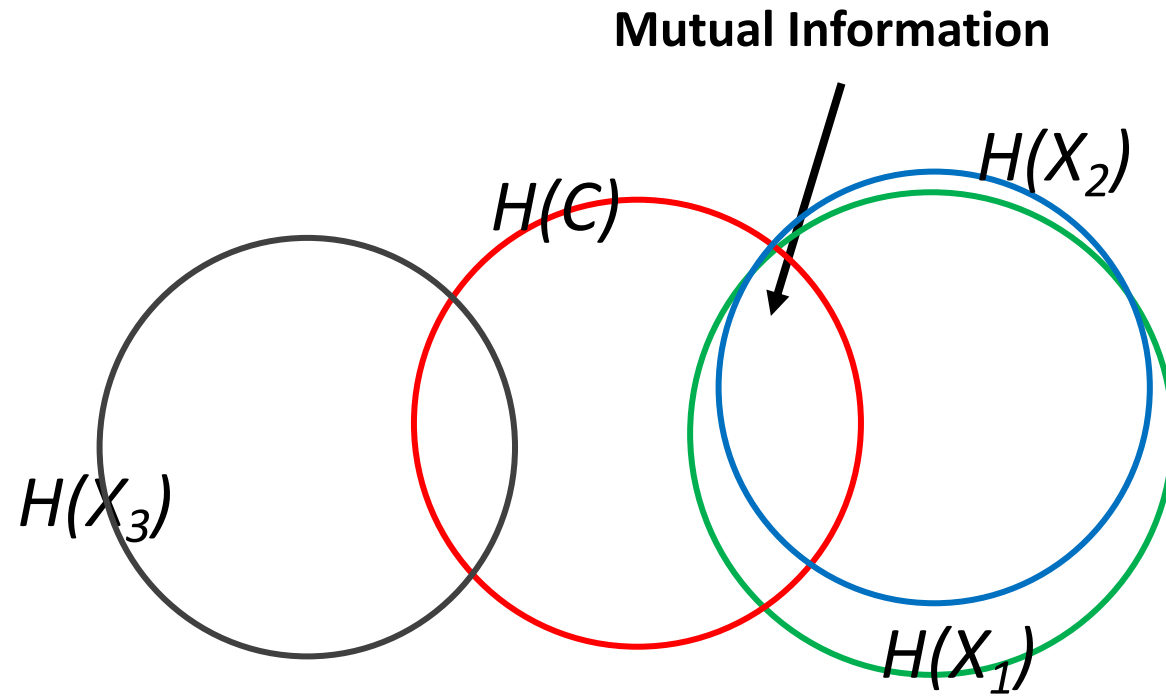
$$NI(F; C) = \frac{I(F; C)}{H(C, F)}$$

$NI(X, Y) = 0$, which means the two variables are totally independent.

$NI(F1, C)$	1
$NI(F2, C)$	1
$NI(F3, C)$	0.1242

F1	F2	F3	C
11	64	1	2
8	1	1	1
11	64	0	2
14	729	1	3
17	100	0	4
11	64	1	2
8	1	0	1
11	64	1	2

Limitations of NIG



More MI- Based feature selection methods

To tackle this problem new methods have been proposed for selecting relevant features, which are non-redundant with respect to each other.

The aim is to maximise the mutual information between the selected features and the class labels $I(S, C)$

$$I(S; C) = I(f_1, f_2, \dots, f_k; C)$$

Where $S = (f_1, f_2, f_3, \dots, f_k)$ is the selected subset, C is the class label.

Joint Mutual Information Maximisation (JMIM) method

This method employs Joint Mutual Information (JMI) and the maximum of the minimum approach, which should choose the most relevant features.

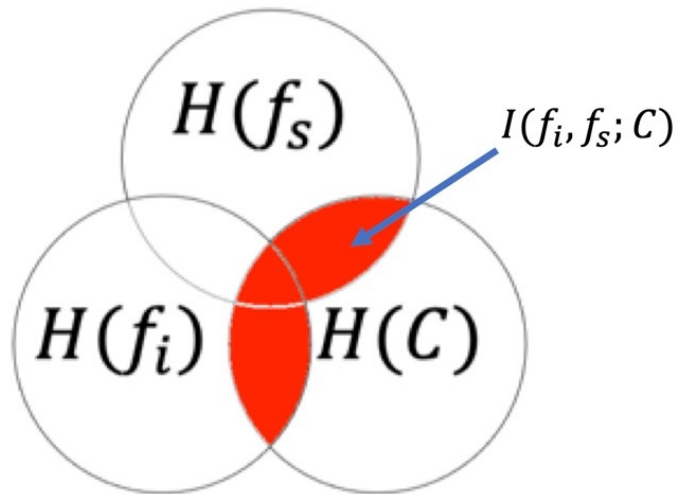
Definition1. (Feature relevance). Feature f_i is more relevant to the class label C than feature f_j in the context of the already selected subset S when $I(f_i, S; C) > I(f_j, S; C)$.

Definition2. Candidate feature f_i is redundant to the selected features within the subset S if f_i does not share new information with the class C .

JMIM

The features are selected by JMIM according to the following criterion:

$$f_{JMIM} = \arg \max_{f_i \in F-S} \left(\min_{f_s \in S} (I(f_i, f_s; C)) \right)$$



JMIM Algorithm

Algorithm 1. Forward greedy search.

1. (Initialisation) Set $F \leftarrow$ "initial set of n features"; $S \leftarrow$ "empty set."
 2. (Computation of the MI with the output class) For $\forall f_i \in F$ compute $I(C; f_i)$.
 3. (Choice of the first feature) Find a feature f_i that maximises $I(C; f_i)$; set $F \leftarrow F \setminus \{f_i\}$; set $S \leftarrow \{f_i\}$.
 4. (Greedy selection) Repeat until $|S| = k$: (Selection of the next feature) Choose the feature $f_i = \arg \max_{f_i \in F \setminus S} (\min_{f_s \in S} (I(f_i, f_s; C)))$; set $F \leftarrow F \setminus \{f_i\}$; set $S \leftarrow S \cup \{f_i\}$.
 5. (Output) Output the set S with the selected features.
-

Practical example

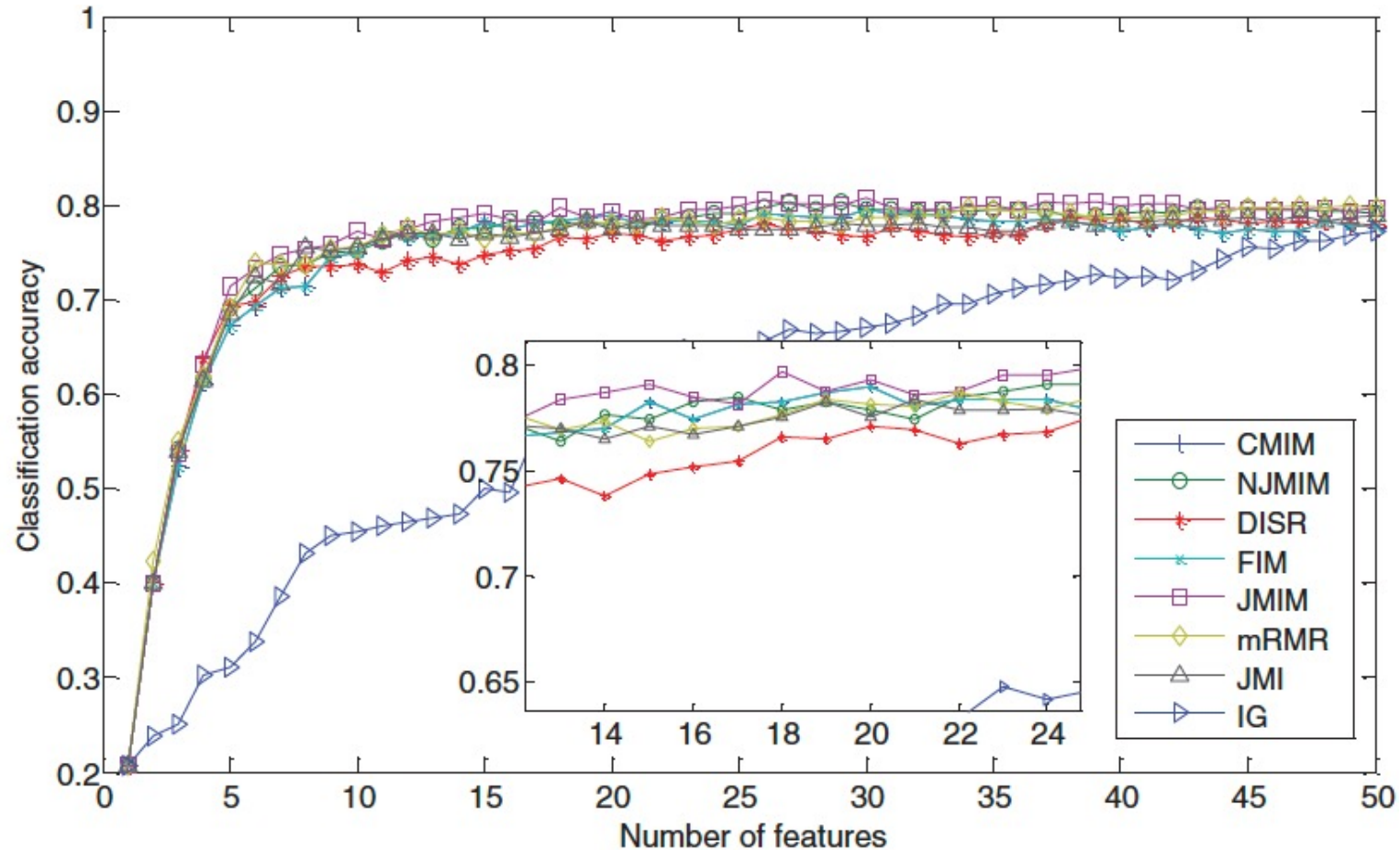


Fig. 9. Average classification accuracy achieved with the *libra* movement dataset.

Feature extraction

Feature extraction methods transform existing features into a new feature space of lower dimensionality.

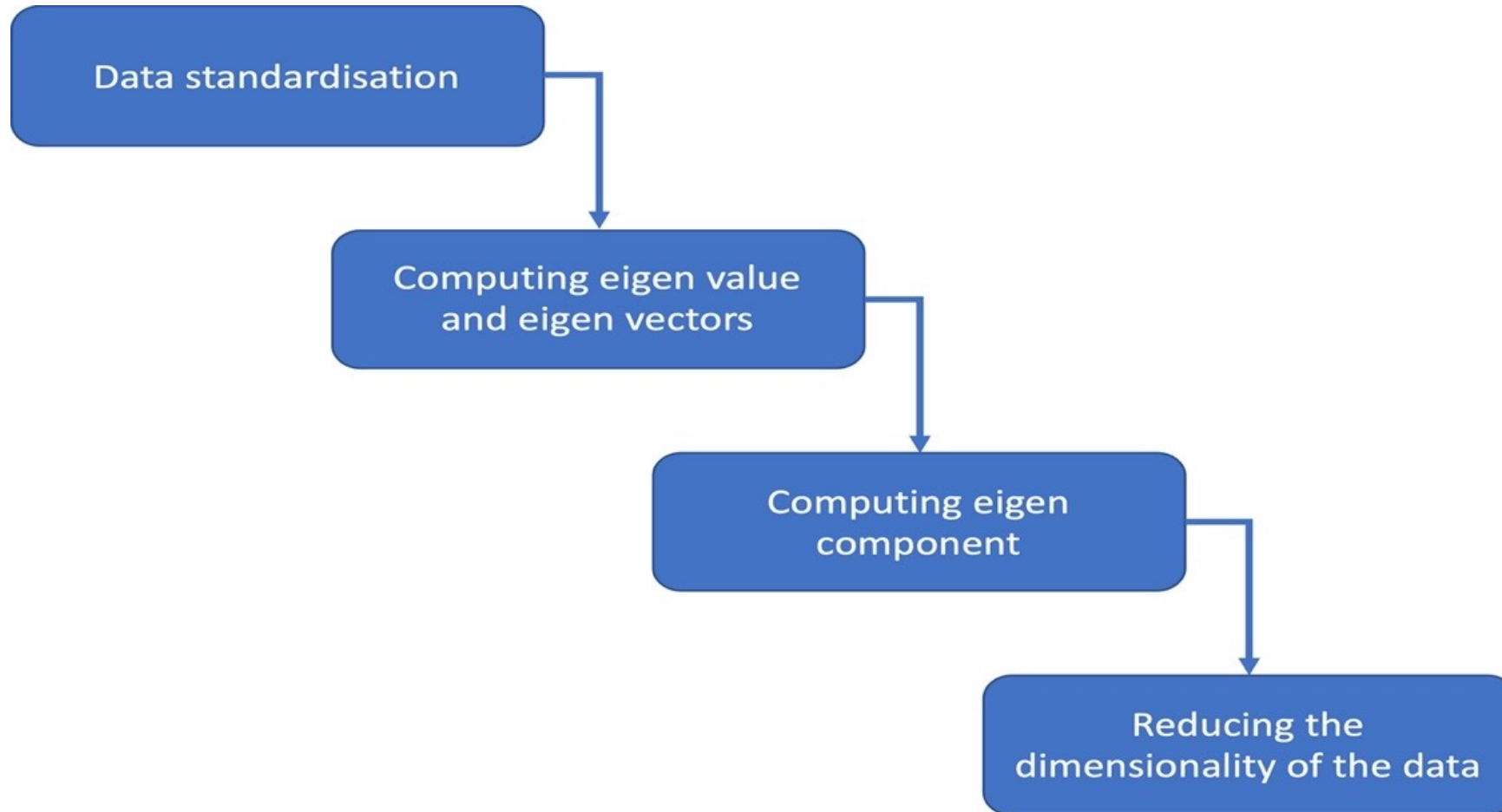
Principal Component Analysis (PCA)

PCA orthogonal transformation method that transforms the original features into a new feature which is called eigen components.

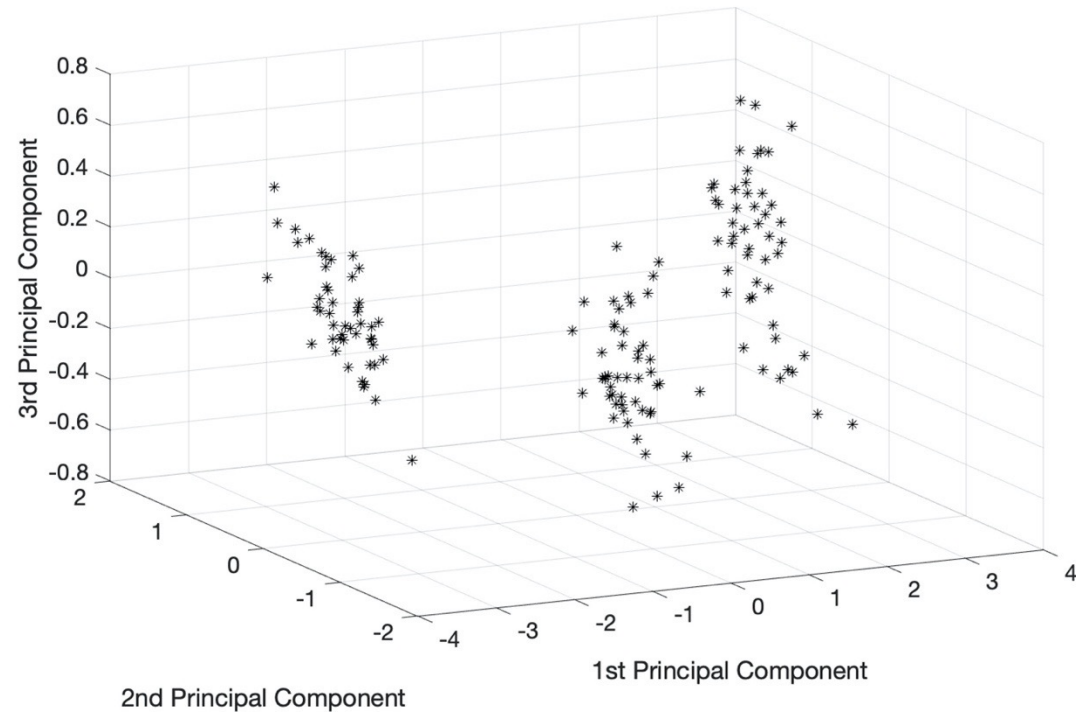
The original features are project into coordinate system that shows the highest variation in the data.

The highest variation will be along the first eigen vector, the second highest will be along the second eigen vector, and so on.

Principal Component Analysis (PCA) dimensionality reduction pipeline:



IRIS dataset



For more information

- **Elements of information theory**, John Wiley & Sons, New York (2006)
- Bennasar, M., Hicks, Y. and Setchi, R., 2015. **Feature selection using joint mutual information maximisation**. *Expert Systems with Applications*, 42(22), pp.8520-8532.