UNIVERSITY OF EDINBURGH
Business School

Predictive Analytics and Modelling of Data

CMSE11428 (2020-2021)

Dr Xuefei Lu
The University of Edinburgh Business School

Unsupervised Learning

Supervised VS unsupervised

Predictive Model: $y = f(X_1, X_2, \dots, X_p)$

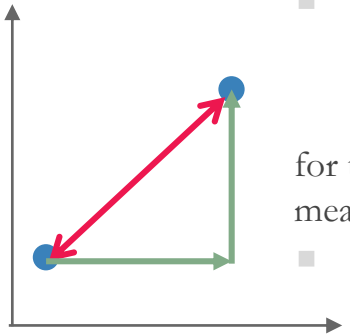
- The majority of practical machine learning uses supervised learning.
- *Supervised learning is where you have input data (**X**) and output data (**y**) and you use an algorithm to learn the mapping function **f** from the input to the output.*
- Classification when y is categorical
  Regression when y is continuous
- E.g. linear regression, random forest, neural network, support vector machine…

- *Unsupervised learning is where you only have input data (**X**) and NO corresponding output data.*
- The goal for unsupervised learning is to model the underlying structure/pattern in the data.
- Clustering to discover the inherent groupings, e.g. customer segmentation
- Association to discover rules that describe large portions of your data, e.g. people that buy A also tend to buy B.
- K-means, Apriori…

Distance functions and clustering

Clustering identifies clusters within the dataset of similar items.

We use distance measures to create clusters that can quantify how far away observations are from each other.



- **The Euclidean distance** (L-2 norm)

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\Sigma_{i=1}^{d}(x_i - y_i)^2}$$

for two observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)$ of dim d. In 2-dim case, it measures the straight line connecting two points.

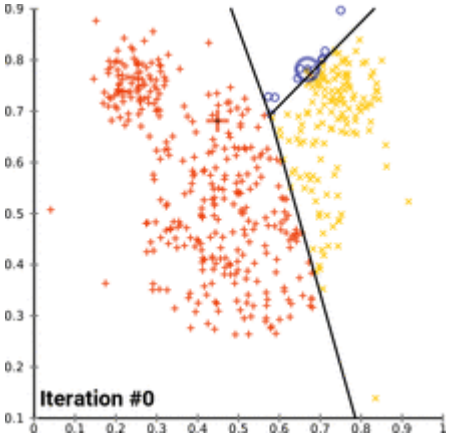
- **The Manhattan distance**:

$$d_M(\mathbf{x}, \mathbf{y}) = \Sigma_{i=1}^{d}|x_i - y_i|$$

This distance sums up the straight distances of the dimensions along their axis.

- Clustering methods can only apply to numerical variables.
- For Categorical, convert to numeric data/use *similarity through overlap* (how many values overlap between two observations) → not recommended

K-means clustering algorithm



https://en.wikipedia.org/wiki/K-means_clustering

1. It starts initially by randomly selecting k different observations to act as centroids for the k clusters we want to find.
2. Next, assign each observation to the cluster which the centroid is the closest to the observation in terms of the Euclidean distance.
3. Then, re-calculates the centroid of the cluster by taking the multidimensional mean of all observations in the current cluster.
4. Repeat Step 2 and Step 3, until maximum iterations has reached, or the clusters don't change anymore.

K-means clustering algorithm

- We cannot know how many clusters k in priori.
  ➔ try different values
- In practice, it is computationally fast, but may suffer the 'curse of dimensionality' when the number of dimensions are high.
- The algorithm does not guarantee convergence to the global optimum.
- The result may depend on the initial clusters.
  Can be either be random or based on dissimilarity to achieve faster convergence.
- Variants of the k-means algorithm: k-medians algorithm, algorithm based on Manhattan distance,...

Quiz

In which of the of following instances does the Euclidean distance suffers from the curse of dimensionality?

- A. When there is a high number of dimensions.
- B. When there is a low number of dimensions.
- C. It is unrelated to the number of dimensions, but is related with the size of the dataset.
- D. It does not suffer from the curse of dimensionality.

Quiz

A bad initialisation of K-means can lead to:

- A. More clusters.
- B. Fewer clusters.
- C. Longer running times.
- D. All of the above.

- Study file
  10 - clustering_data.ipynb

Activity: Generating and Clustering data

UNIVERSITY OF EDINBURGH
Business School