

MS4S10 Machine Learning and Decision Making

Week 4

Moizzah Asif

moizzah.asif@southwales.ac.uk

J418

1

Know thy module



Week 1 – 4
Moizzah Asif

- 27-11-2020 – Basics of Machine Learning; The machine learning process; Data collection & preprocessing
- 04-12-2020 – Supervised Learning: classification, regression, optimisation, model selection and generalisation, parametric and non-parametric learning, Decision Trees
- 11-12-2020 – Supervised Learning: Probabilistic learning, Bayes Learning, Naïve Bayes classifiers, Ensemble Learning, Random Forest, Support Vector Machines, Kernel functions, Hyper-parameter optimisation
- 08-01-2021 – Unsupervised Learning: clustering and dimensionality reduction

Course work 1 – 50% weightage

2

Unsupervised Learning

Introduction

Correct answer is not provided:
Input data has no known label/ result

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 - i. Data Preparation
 - ii. Algorithm Selection

Hence, data is not split into test and train sets.

A model is generated from analysing the structures present in the input data.

3

Unsupervised Learning

Introduction

The model may be generated by:
a rule,
or by a mathematic process

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 - i. Data Preparation
 - ii. Algorithm Selection

The model may be generated to:
reduce redundancy,
or to organise data by characteristic shared in common.

There are a variety of algorithms that utilise this learning approach, such as:
clustering,
or dimensionality reduction.

4

Unsupervised Learning - Clustering

Introduction

Clustering is the organisation of data into similarity groups.

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 - i. Data Preparation
 - ii. Algorithm Selection

A cluster has data instances which are similar to each other and dissimilar to those in other cluster

5

Unsupervised Learning - Clustering

Clustering Techniques – K-means

Introduced by MacQueen in 1967

Is a *partitional* clustering algorithm

Partitions the data into ***k* cluster**

In a K-means clustering algorithm:

there are ***k*** clusters

each cluster has a centre, called ***centroid***

the value of ***k*** is selected by the user

6

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Clustering Techniques – K-means PseudoCode

Given k ,

1. Choose k (random) data points (seeds) to be the initial centroids
2. Assign each data point to the closest centroid
3. Re-compute the centroids using the current cluster memberships
4. If a convergence criterion is not met, repeat steps 2 and 3

7

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

7

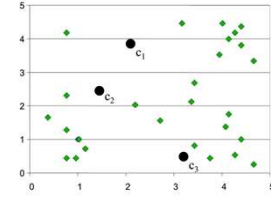
University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Clustering Techniques – K-means Example

a two dimensional dataset

randomly initialise centroids for $k = 3$



8

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

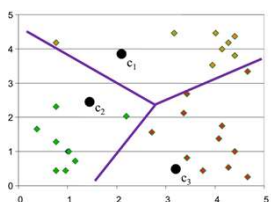
8

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Clustering Techniques – K-means Example

Determine cluster membership for each instance



9

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

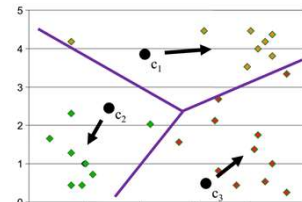
9

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Clustering Techniques – K-means Example

Re-estimate cluster centres



10

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

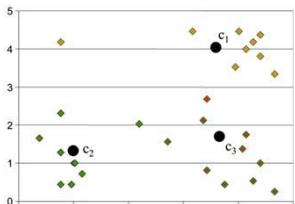
10

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Clustering Techniques – K-means Example

First iteration looks as follows



11

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

11

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Clustering Techniques – K-means Example

Assess the cluster based on the stopping criterion

Possible stopping criteria

- No or minimum reassignments of data points to different clusters
- No or minimum change of centroids
- A user defined tolerance of similarity is reached

12

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

12

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

Clustering Techniques – K-means

What is similarity? Opposite of distance

How to measure it? Minimising the sum of squared error (SSE)

$$SSE = \sum_{j=1}^k \sum_{x \in C_j} d(x, m_j)^2$$

13

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

13

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

K-means – strength

Simple: easy to understand and to implement

K-means – weakness

The algorithm is only applicable if the mean is defined.

For categorical data, k-mode - the centroid is represented by most frequent values

The user needs to specify k

The algorithm is sensitive to outliers

The algorithm is sensitive to initial seeds

14

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

14

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

K-means – weakness

The algorithm is only applicable if the mean is defined.


For categorical data, k-mode - the centroid is represented by most frequent values

The user needs to specify k

The algorithm is sensitive to outliers

The algorithm is sensitive to initial seeds

algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres)



15

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

15

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

K-means – Dealing with outliers

Remove some data points that are much further away from the centroids than other data points

better to monitor these possible outliers over a few iterations and then decide to remove them

Can choose a small subset of the data points

16

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

16

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

K-means – finding optimal number of clusters

Intuition – when there isn't any significant decrease in within cluster SSE

Can be visually represented using **elbow method**

17

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

17

University of South Wales
Prifysgol De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

K-means - Quantifying the clusters quality

Silhouette analysis – find silhouette co-efficients

1. Calculate the cluster cohesion a^i as the average distance between a sample x and all other points in the same cluster
2. Calculate the cluster separation b^i from the next closest cluster as the average distance between the sample x and all the samples in the nearest cluster
3. Calculate silhouette for that sample as shown below

$$s^{(i)} = \frac{b^{(i)} - a^{(i)}}{\max\{b^{(i)}, a^{(i)}\}}$$

18

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

18

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Silhouette analysis – silhouette coefficients

ranges from -1 to 1

Coefficient is 0 when separation and cohesion are equal (worst case)

Coefficient is 1 when separation (b) is greater than cohesion (a) (ideal case)

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

19 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

19

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Hierarchical Clustering

K-means is a flat clustering algorithm

Whereas hierarchical clustering maintains hierarchies between clusters

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

20 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

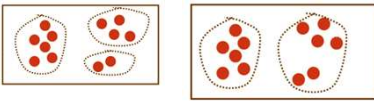
20

University of South Wales
Prifysgol De Cymru


Unsupervised Learning - Clustering

Hierarchical Clustering

K-means is a flat clustering algorithm



Whereas hierarchical clustering maintains hierarchies between clusters



1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

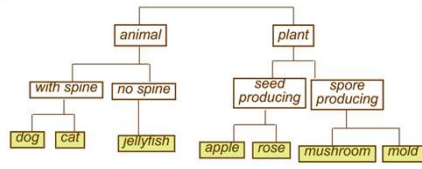
21 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

21

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Hierarchical Clustering



1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

22 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

22

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Hierarchical Clustering - Divisive

Develops hierarchy of clusters taking a top down approach

Splits the root into a set of child clusters.

Each child cluster is recursively divided further

stops when only singleton clusters of individual data points remain, i.e., each cluster with only a single point

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

23 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

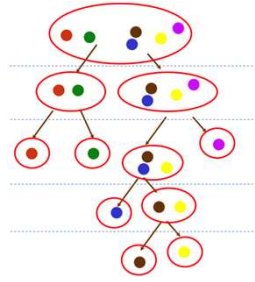
23

University of South Wales
Prifysgol De Cymru

Unsupervised Learning - Clustering

Hierarchical Clustering - Divisive

Can be developed using any flat clustering algorithm



1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

24 Moizzah Asif - Machine Learning and Decision Making © University of South Wales

24

University of South Wales Prifysgol De Cymru

Unsupervised Learning - Clustering

Hierarchical Clustering - Agglomerative

Develops hierarchy of clusters taking a bottom up approach

Each instance is considered as a cluster

Each child cluster is recursively built up further by merging with most similar (nearest) clusters

stops when all data points are merged into a single cluster

25

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

25

University of South Wales Prifysgol De Cymru

Unsupervised Learning - Clustering

Hierarchical Clustering – Agglomerative

There are four common ways to find nearest clusters:

minimum distance: $d_{\min}(D_i, D_j) = \min_{x \in D_i, y \in D_j} |x - y|$

maximum distance: $d_{\max}(D_i, D_j) = \max_{x \in D_i, y \in D_j} |x - y|$

average distance: $d_{\text{avg}}(D_i, D_j) = \frac{1}{n_i n_j} \sum_{x \in D_i} \sum_{y \in D_j} |x - y|$

mean distance: $d_{\text{mean}}(D_i, D_j) = |\mu_i - \mu_j|$

26

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

26

University of South Wales Prifysgol De Cymru

Unsupervised Learning - Clustering

Hierarchical Clustering – Agglomerative

Single link agglomerative cluster uses:

minimum distance: $d_{\min}(D_i, D_j) = \min_{x \in D_i, y \in D_j} |x - y|$

Generates minimum spanning tree

Hence encourages elongated clusters

Sensitive to noise

27

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

27

University of South Wales Prifysgol De Cymru

Unsupervised Learning - Clustering

Hierarchical Clustering – Agglomerative

complete link agglomerative cluster uses:

maximum distance: $d_{\max}(D_i, D_j) = \max_{x \in D_i, y \in D_j} |x - y|$

encourages compact clusters

Does not work well when elongated clusters are inherent to the data

Sensitive to noise

28

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

28

University of South Wales Prifysgol De Cymru

Unsupervised Learning - Clustering

Hard vs Soft Clustering

Hard Clustering
Each sample is assigned to one cluster

Soft Clustering
Assigns a sample to one or more cluster
Also called fuzzy clustering

29

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

29

University of South Wales Prifysgol De Cymru

Unsupervised Learning - Clustering

Since the task of clustering is subjective, there are a range of different clustering algorithms that follow different sets of rules for identifying **similarity** amongst data points:

Connectivity Models (Hierarchical clustering),

Centroid Models (K-means),

Distribution Models ,

Density Models

30

Molizah Asif - Machine Learning and Decision Making

© University of South Wales

30

University of
South Wales
Prifysgol
De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

Clustering Rules – Connectivity Models

These models are based on the assumption that the data points closer in data space exhibit more similarity to each other than the data points located further away.

There are different methods for calculating distances:

- Single-linkage – minimum of object distances,
- Complete-linkage – maximum of object distances,
- Average-linkage – Unweighted or weighted arithmetic mean.

31
Molizah Asif - Machine Learning and Decision Making
© University of South Wales

31

University of
South Wales
Prifysgol
De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

Clustering Rules – Centroid Models -I

These models are iterative clustering algorithms that assume similarity is related to the closeness of a data point to the centroid of the cluster.

In these models, the number of clusters have to be pre-determined.

This makes it important to have prior knowledge of the dataset.

32
Molizah Asif - Machine Learning and Decision Making
© University of South Wales

32

University of
South Wales
Prifysgol
De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

Clustering Rules – Distribution Models

These clustering models are closely related to statistics and are based on how probable is it that all data points in a cluster belong to the same distribution.

For example - normal, Gaussian distribution.

The models often suffer from overfitting.

33
Molizah Asif - Machine Learning and Decision Making
© University of South Wales

33

University of
South Wales
Prifysgol
De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

Clustering Rules – Density Models

These clustering models search the data space for areas of varied density of data points.

Different density regions are sectioned off and data points within these regions are assigned to the same cluster.

34
Molizah Asif - Machine Learning and Decision Making
© University of South Wales

34

University of
South Wales
Prifysgol
De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

Density Models - DBSCAN

Does not need a priori number of clusters

Can capture cluster of complex shapes

Can also identify points which are not part of any cluster(potential noise)

Finds sample that are in crowded regions(dense regions)

The idea is clusters from dense regions of data are separated by regions that are relatively empty

35
Molizah Asif - Machine Learning and Decision Making
© University of South Wales

35

University of
South Wales
Prifysgol
De Cymru

1. Basics of Machine Learning (ML)
2. The Machine Learning Process
 i. Data Preparation
 ii. Algorithm Selection

Unsupervised Learning - Clustering

Density Models – DBSCAN

How it works

Core samples – instances in a dense region

Parameters of DBSCAN

$\min_samples$
 eps

atleast $\min_samples$ many samples with in a distance eps to a given sample x , then x is considered as core sample.

36
Molizah Asif - Machine Learning and Decision Making
© University of South Wales

36

University of
South Wales
Prifysgol
De Cymru

Unsupervised Learning - Clustering

Density Models – DBSCAN

How it works

1. Basics of Machine Learning (ML)

2. The Machine Learning Process

I. Data Preparation

II. Algorithm Selection

Picks an arbitrary point

Finds all points with in *eps* distance of that point

If there are less than *min_samples points* with in *eps* distance of this point, it is labelled as **noise**

if there are more than *min_samples* datapoints with in *eps* then the point is labelled as core sample, and assigned a new cluster label

all neighbouring data points with in *eps* are visited – unassigned core samples are assigned to this cluster

these core samples' neighbours are visited, and so on

37
Molazzah Asif - Machine Learning and Decision Making
© University of South Wales

37

University of
South Wales
Prifysgol
De Cymru

Unsupervised Learning - Clustering

Density Models – DBSCAN

How it works

1. Basics of Machine Learning (ML)

2. The Machine Learning Process

I. Data Preparation

II. Algorithm Selection

The cluster grows until there are no core samples with in *eps* distance of cluster

Another point which has not been visited until now is picked and the whole process repeats

38
Molazzah Asif - Machine Learning and Decision Making
© University of South Wales

38