7. K-Means (Unsupervised Learning)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

MACHINE LEARNING OVERVIEW

A set of methods for instructing the computer to learn from the data on its own

- Somewhat related to conventional data analytics frameworks, e.g. Statistics, but also independent
- Related to A.I. but not the same
  - Comprises of Supervised, Unsupervised, and Semi-supervised (Reinforcement) Learning

This class's focus: unsupervised learning, particularly clustering

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

CLUSTERING

What is it?

A method for organizing data into groups

A kind of unsupervised learning (=> no labels a priori)

An effective way to obtain potential labels

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

CLUSTERING

Why is it useful?

Data is oftentimes unlabelled

Clustering enables for a better understanding of the structure of a dataset

An essential part of data exploration

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Important Considerations in Clustering

- Requires numeric (or Binary) data
- All features need to be of the same scale
- Doesn't work very well with very large number of features (curse of dimensionality)
- Pay attention to the distance metric
  - Euclidean distance, Manhattan distance, etc.
- Often it is best to perform visualization before and after the clustering process

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Some Applications of Clustering

- Figure out the main categories of the users of an app, based on some data we gather from them
- Organize a corpus into topics based on the most important words in the documents
- Explore relationships among variables
- Gain a better understanding of our data in general and insights on how to frame a data science problem

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

CLUSTERING

How do we actually perform clustering?

Clustering is an NP problem => no exact solution

Various clustering algorithms out there, mainly stochastic in nature (i.e. not deterministic)

Most popular (and simple) one is K-means

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

K-MEANS ALGORITHM OVERVIEW

Inputs:

a matrix X comprising of N data points across m dimensions (features)

a parameter k related to the expected number of clusters

a parameter th related to when the algorithm converges

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

K-MEANS ALGORITHM

Outputs:

a vector Y comprising of N elements, corresponding to the assigned cluster of each data point in X

a matrix C comprising of k rows and m columns, each row corresponding to the center of a cluster (aka centroid)

– Every time K-means is run, it yields a somewhat different result

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

K-MEANS ALGORITHM

Algorithm pseudocode

1. Initialize centroids randomly => $\mu_1, \mu_2, \ldots, \mu_k$
2. Do until convergence*
3. For every i, set c(i) = arg min ($|x_{min} - \mu_j|^2$), for all i = 1…N
4. For every j, set $\mu_j = \Sigma\{c_i = j\}x_i \ / \ \Sigma\{c_i = j\}$, for all j = 1…k
5. Repeat steps 2-4

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

K-means and Similarity

The idea behind K-means is to split the dataset into clusters so that the data points in any given cluster are

- all similar to each other
- dissimilar to those of the other clusters

Similarity can be measured in many ways

Most common way: some reverse function of the distance (i.e. small distance = large similarity)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Distance metrics used

Vary significantly from one another

Common distance metrics in clustering:

- Euclidean distance $d(x,y) = (\Sigma[(x-y)^2])^{0.5}$
- Manhattan distance $d(x,y) = \Sigma[|x-y|]$
- "sup" distance $d(x,y) = \max(|xi - yi|)$, for $i = 1...m$

In the case of binary features, Manhattan distance is often referred to as Hamming distance

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Parameter selection

k (number of clusters)

- Has to be an integer
- Needs to correspond to the geometry of the dataset, for best results
- Typically between 2 and 10, though it could be higher for some applications

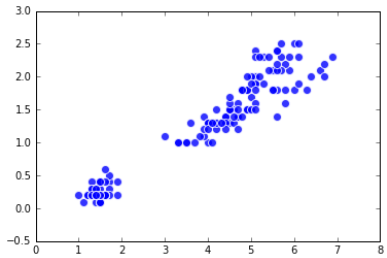
Parameter selection

th (minimum shift in centroids)

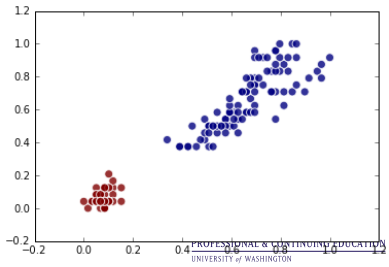
- Has to be a float number (e.g. 0.00001)
- If it's very small => longer time for k-means to terminate
- If it's too large => results may be unreliable
- Typical values: 0.000001 to 0.001

K-means example

Plot of original dataset (Iris flowers, features 3 and 4)



Plot of dataset after been clustered using K-means



Summary

>Clustering provides labels on the data
  –Based on similarity of the data
>Perform EDA before and after clustering
>K-means clustering is affected by:
  –Number of centroids
  –Distance metric used
  –Convergence (termination) parameter