SEG4135 - Lecture 7

Big Data Analytics

Big Data Analytics Approaches

Will fuel the industry for the next decade
Big data is not mining on a static database

Fundamental characteristics of Big Data

Volume
- There's lots of it

Velocity
- It grows really quick

Variety
- It may come in all forms.

Veracity
- What it says

Value
- How we can capitalize on it

Clustering of Big Data

Also known as unsupervised learning. The data is not labeled, we don't know how many unique patterns there are.

Useful for
- Social network data
- Election health record
- Sensor data to group similar or related faults in a machine
- Clustering market research data to group similar customers
- Clickstream data to group similar users.

Put things into similar groups, but we don't know what these groups mean.

K-mean clustering algorithms

You have a group of data points and we want to put the data points into these clusters.

We have a core for each cluster, each core has a centroid.

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }

If you don't have a data point that fits any of your clusters, it will be put into a cluster that doesn't represent it.

DBScan

Density-based spatial clustering of applications with noise (*DBSCAN*)

Based on "neighbourhoods" of clusters, where for every point p in a cluster C there is a point q in c to that p is inside of the neighbourhood of q.

Classification of Big Data

Also known as supervised learning. You have a training set and know what the data will look like.

Classification is the process of categorizing objects into predefined categories. We are working with a labeled dataset.

Recommendation Systems