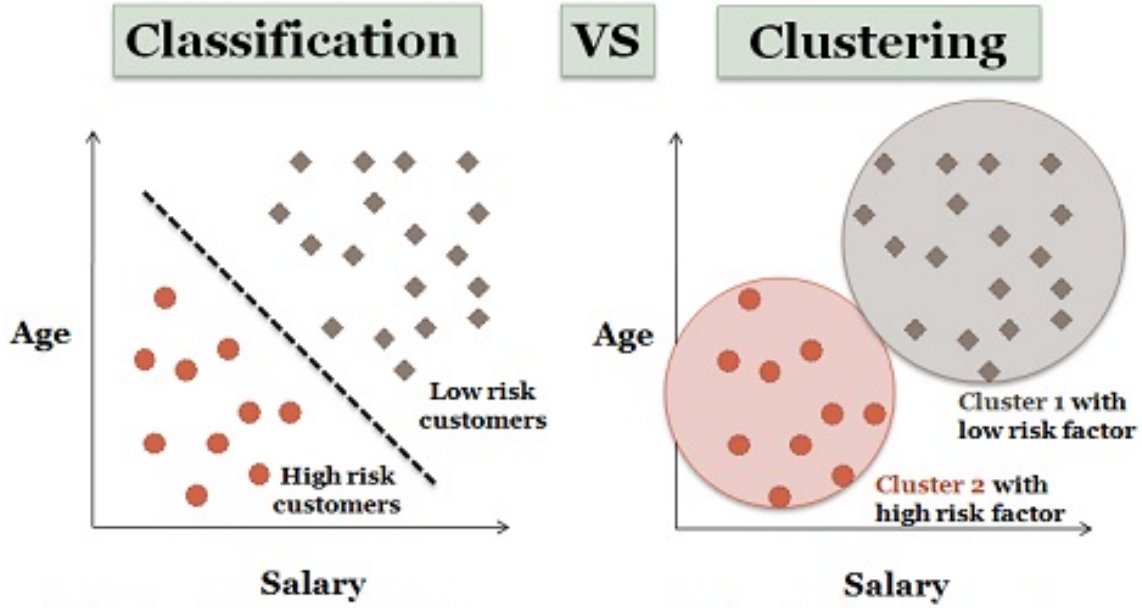
Big Data Analysis

Lecture 12

2019/12/26

Clustering vs. Classification

- Unsupervised vs. supervised learning
- K-Means vs. SVM



Risk classification for the loan payees on the basis of customer salary

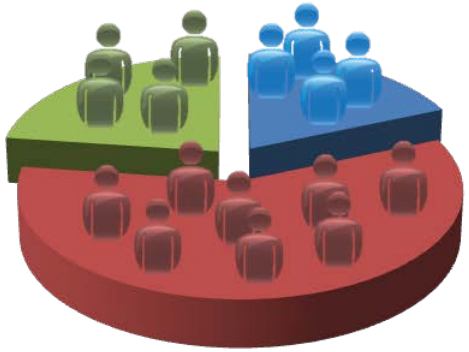
https://techdifferences.com/difference-between-classification-and-clustering.html

K-Means

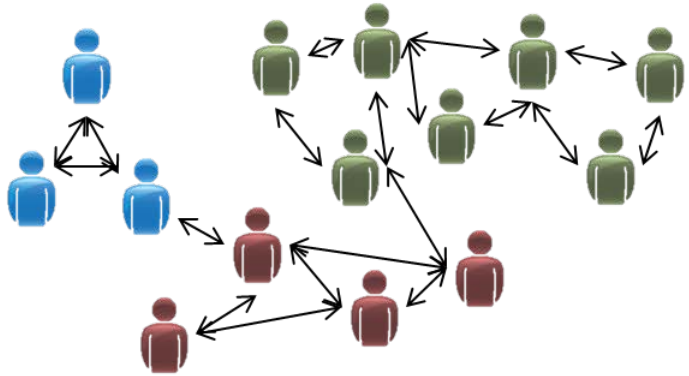
- A short summary from Professor Andrew Ng at http://cs229.stanford.edu/notes/cs229-notes7a.pdf

- Course slides can be found at https://github.com/vkosuri/CourseraMachineLearning/blob/master/home/week-8/lectures/ppt/Lecture13.ppt

Motivation for K-Means



Market segmentation



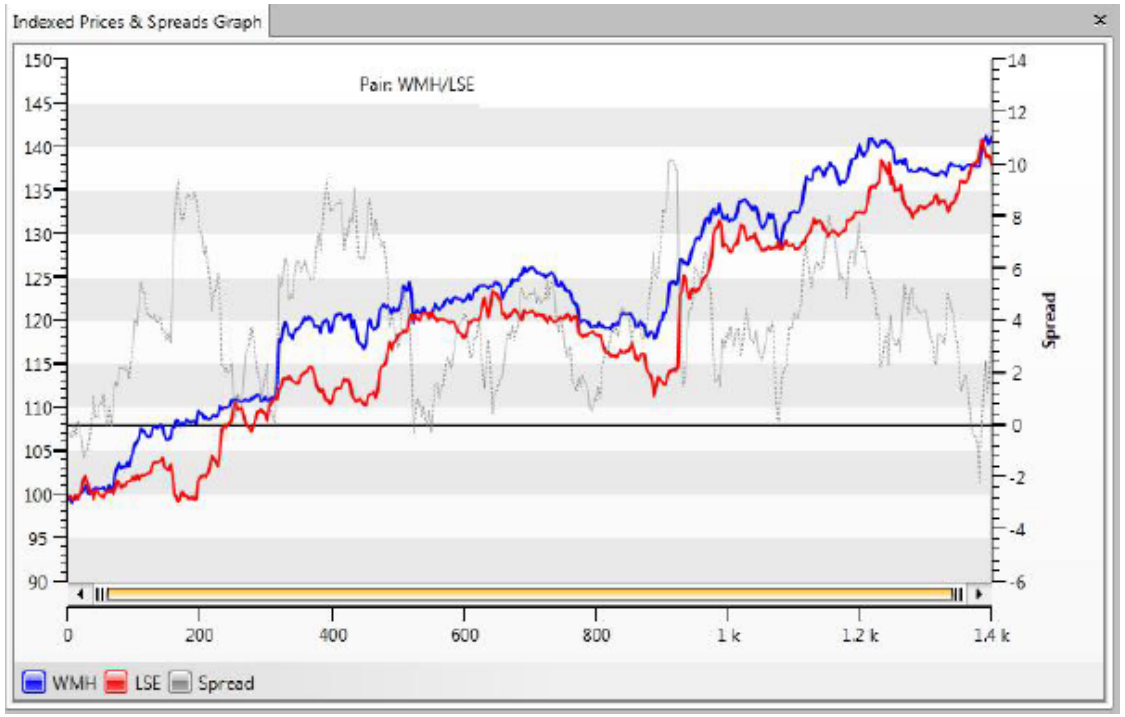
Social network analysis

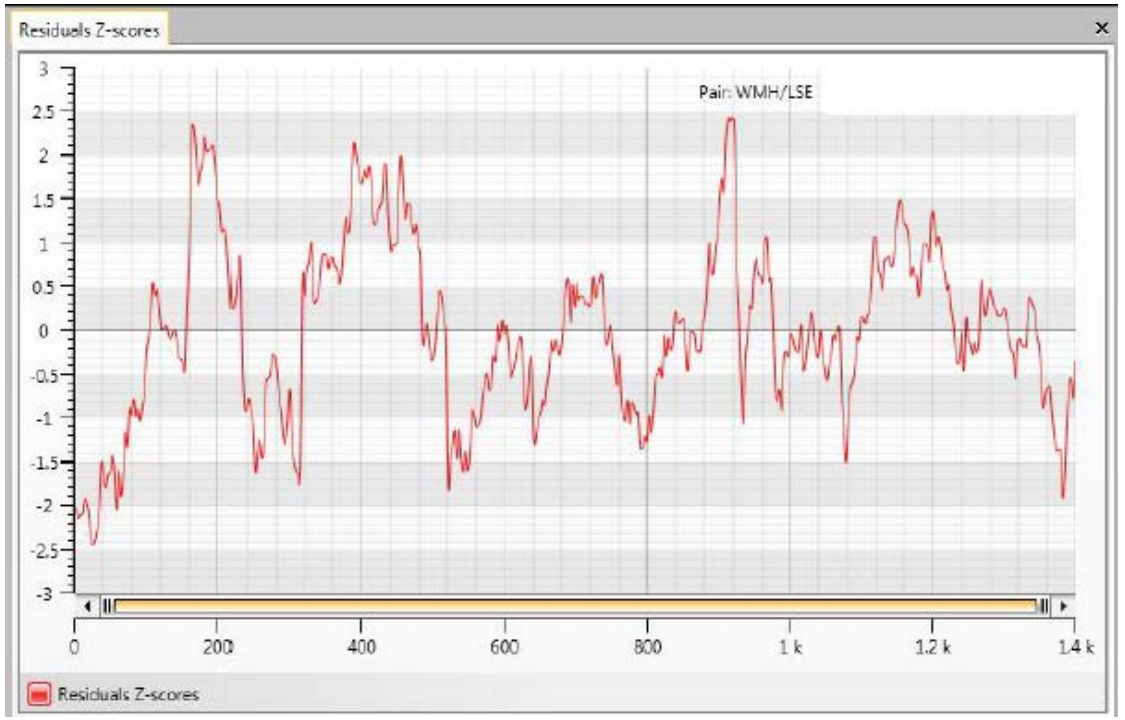
Organize computing clusters

Astronomical data analysis

The above example is borrowed from Professor Andrew Ng's slides.

Stat Arb Strategy





The above example is borrowed from Professor Xiaotian Zhu's slides.

Modeling the Spread

- Correlation vs. Co-Integration
- Co-Integration steps
  - Estimate a co-integrating relationship using the Engle-Granger two-step method or the Johansen procedure
  - Model the residuals with an Ornstein-Uhlenbeck process,

$$dx_t = \theta(\mu - x_t)\,dt + \sigma\,dW_t$$

  - Calibrate the mean reversion speed and the boundaries of normal range of residuals
- Trades are triggered when residuals are out of the bounds

Challenges

- Co-integrating relationship is unstable
- Parameter calibration window size matters
- Usually two steps
  - Pair screening
  - Trading pairs
- Given that China A Share market has around 3800 stocks as of now, how many pairs to check?
- What about 3 or more stocks to be traded at the same time?
- What about including derivatives?