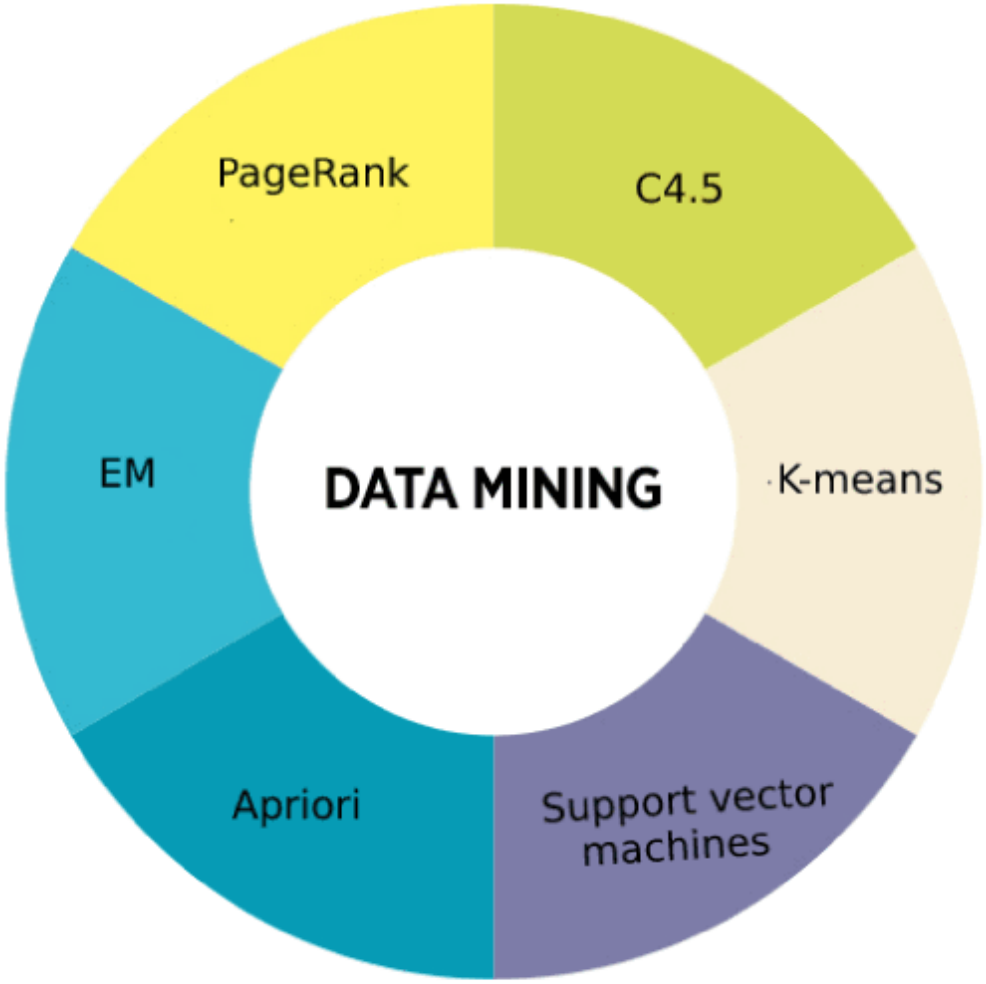
Top Data Mining Algorithms

CSCI 347, Data Mining

TechLeer, 2017



Top Data Mining Algorithms

https://www.techleer.com/articles/438-a-list-of-top-data-mining-algorithms/

Top Data Mining Algorithms

Top data mining algorithms:

1. C4.5 –decision tree learner "landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date"
2. K-means clustering, - "popular cluster analysis technique used for exploring a dataset"
3. Support vector machines for analysis of regression and classification
4. Apriori association rule learner
5. Expectation-maximization , EM, finds the maximum likelihood estimates of parameters, from statistics

Top Data Mining Algorithms - continued

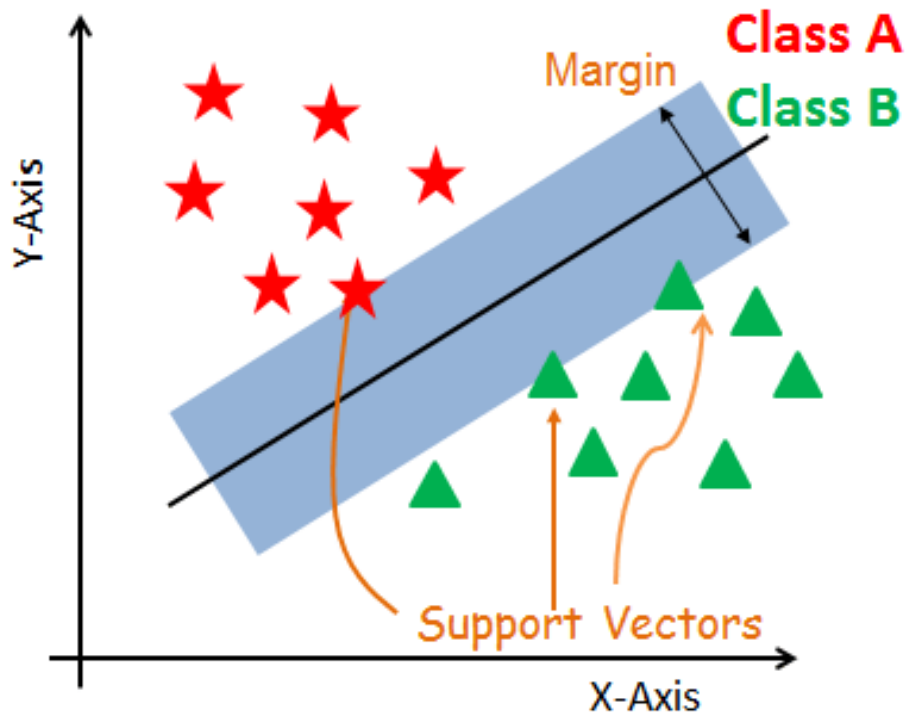
6. PageRank used by Google Search to rank websites
7. AdaptiveBoosting, AdaBoost, meta-algorithm the combines other types of learning algorithm
8. kNN nearest neighbor is among the simplest of all machine learning algorithms
9. Naïve Bayes is a highly scalable simple probabilistic classifier
10. CART decision tree learner that outputs classification or regression trees (like C4.5)

Support Vector Machines (SVM)

From: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

SVM find a maximum marginal hyperplane (MMH) that best divides the dataset into classes.

Support vectors – data points closest to the hyperplane



Support Vector Machines (SVM)

From: https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

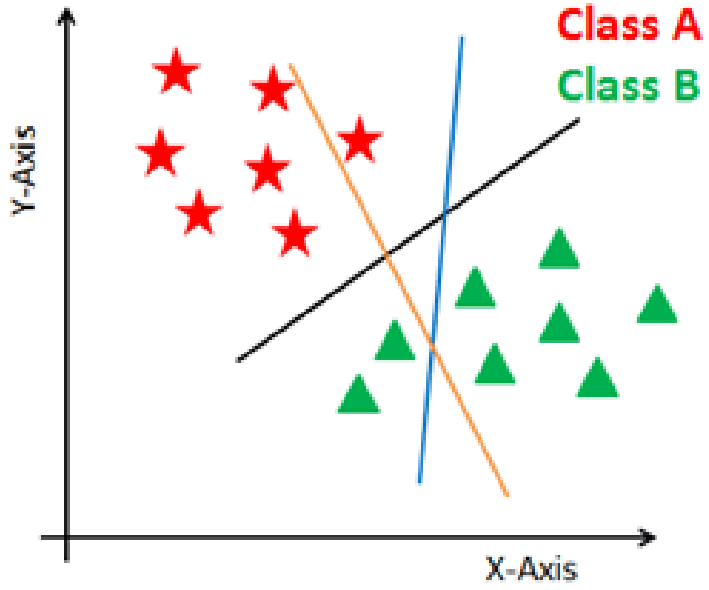
High accuracy compared to logistic regression and decision trees

Used in:

- Face detection (is it a face)
- Handwriting recognition
- Intrusion detection (detect suspicious network activity)
- Classification of emails, news articles and web pages
- Classification of genes

SVM – 2 steps.

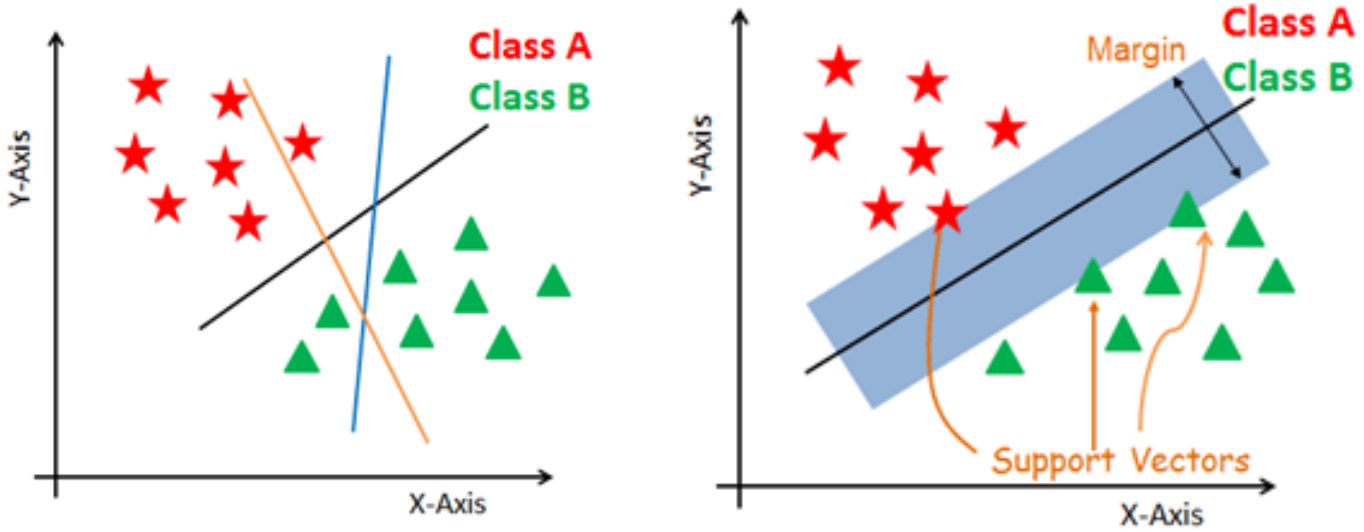
1. Generate hyperplanes which segregates the classes in the best way.



Blue, orange and black hyperplane

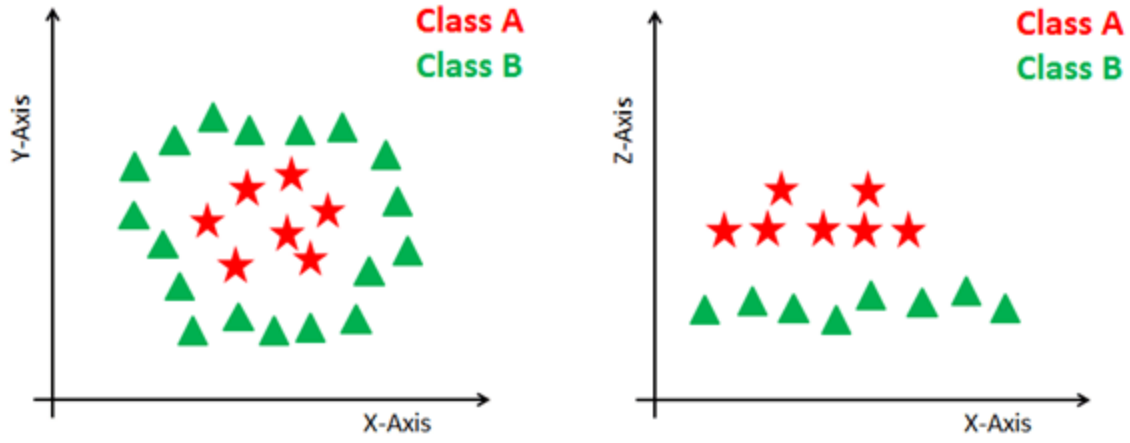
SVM – 2 steps.

2. Select the right hyperplane with the maximum segregation from the either nearest data points



Non-linear and Inseparable Planes

Transform the input space to a higher dimensional space. (Translate points on the left to z-axis which is the squared sum of both x and y, $z=x^2+y^2$.)



Kernel Trick

Transformation, called a kernel trick, takes a low-dimensional input space and transforms it into a higher dimensional space.

It makes a nonseparable problem into a separable problem by adding more dimension to it.

Types of Kernel Tricks

Several types of kernel tricks:

- Linear kernel
  - dot product of any two observations
  - k(x, xi) = sum(x*xi)
- Polynomial kernel
  - Generalized form of the linear kernel
  - $k(x, xi) = 1 + sum(x*xi)^d$ where d is the degree of the polynomial
  - Must input value for 'd'
- Radial basis function kernel
  - Map input space in infinite dimensional space
  - $k(x, xi) = exp(-gamma * sum(x-xi^2))$
  - Must input value for 'gamma' If
  - Large gamma causes overfitting
  - gamma = 0.1 good default

Page Rank

Page Rank – measure, between 0 and 1, of the prestige of a website, named after Larry Page, one of the Google founders

Page rank :

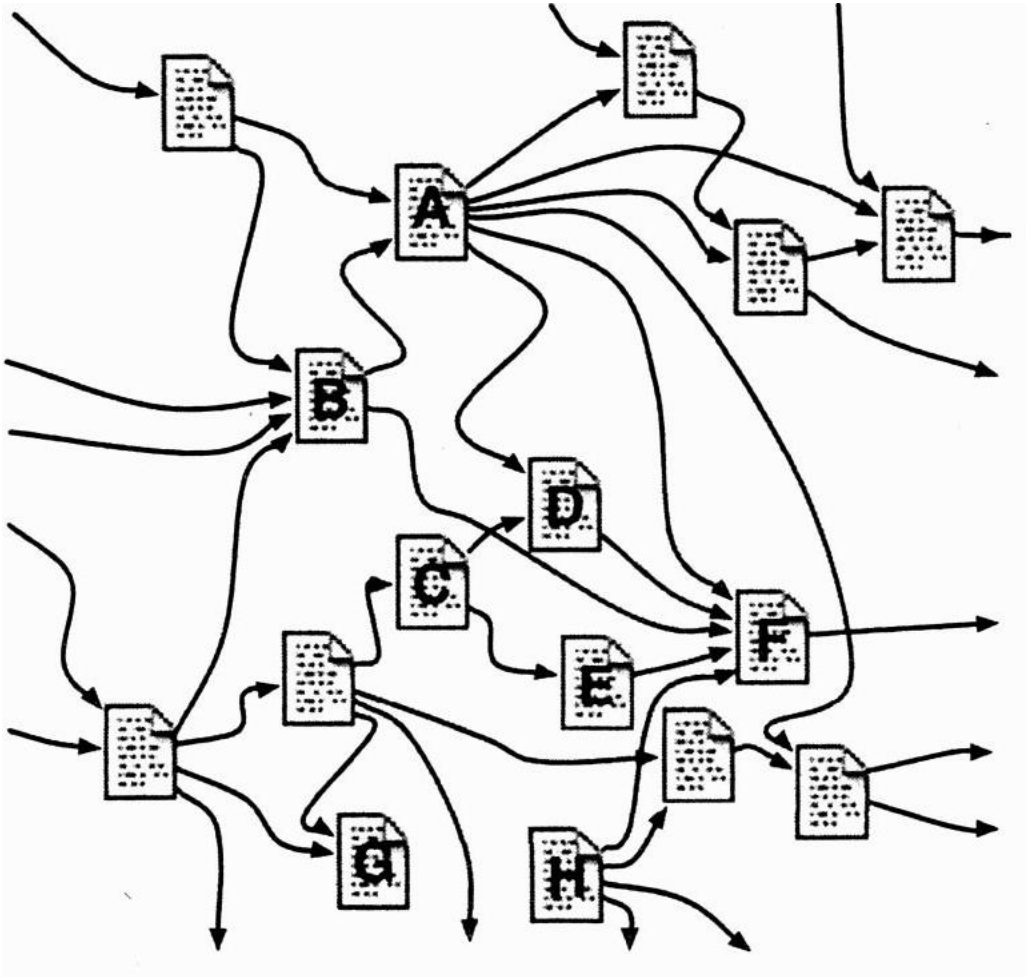
- increases with each in-link (link to the page) by the amount

  page rank of the linking page/

  # out-links from that page

- sum the above to get the page rank

Page Rank



https://www.elsevier.com/books/data-mining/witten/978-0-12-804291-5