CHAPTER 1. INTRODUCTION

By Dr. Benson Lam

Department of Mathematics, Statistics and Insurance

Outline & Content

- What is Machine Learning?
- Supervised Learning
- Unsupervised Learning
- Introduction to Python/Weka

What is Machine Learning?

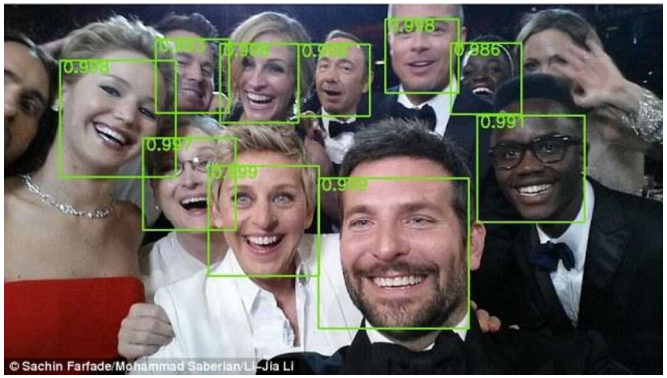
- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.
- Well-posed Learning Problem: A computer program is said to learn from experience with respect to some task and some performance measure.

Supervised Learning

- Data: A set of data records (also called examples, instances or cases) described by
  - *k* attributes: A_1, A_2, … A_k.
  - a class: Each example is labelled with a pre-defined class.
- Goal: To learn a classification model from the data that can be used to predict the classes of new (future, or test) cases/instances.

Supervised Learning – Face detection

- Discriminating human faces from non faces.



Supervised Learning – Face detection

- Face Images.

- Non-face Images

Supervised Learning – Face detection

- Recognition:

Supervised Learning – Face recognition

- Identifying or verifying a person from a digital image.
- Training phase:

Supervised Learning – Face recognition

- Recognition phase:

Mark Zuckerberg?
Obama?
Hiro Nakamura?

Mark Zuckerberg

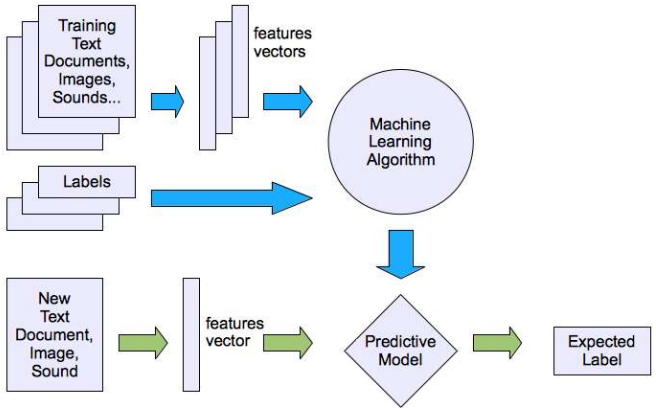
Mark Zuckerberg?
Obama?
Hiro Nakamura?

Obama

Mark Zuckerberg?
Obama?
Hiro Nakamura?

Hiro Nakamura

Supervised Learning



Supervised Learning

- Algorithms
  - Naïve Bayesian classification
  - K-nearest neighbor

Naïve Bayesian Classification

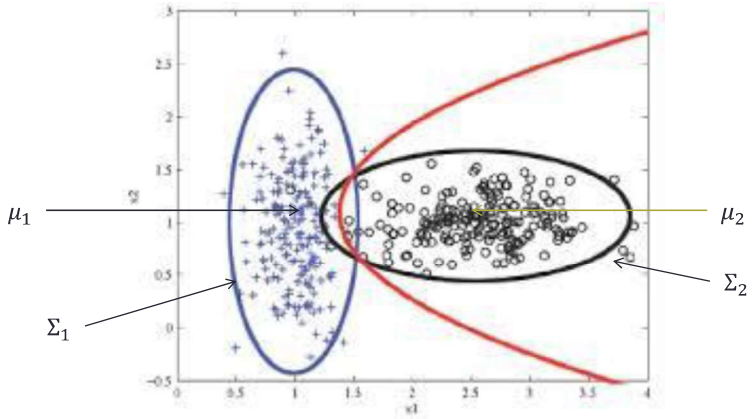
- Gaussian Naïve Bayes
- Assume each group follows a multivariate normal distribution

$$p(x = v|c) = \frac{1}{\sqrt{(2\pi)^k|\Sigma_c|}} e^{-\frac{1}{2}(v-\mu_c)^T\Sigma_c^{-1}(v-\mu_c)}$$

- where μ_c and Σ_c are mean and co-variance matrix of group c

Naïve Bayesian Classification

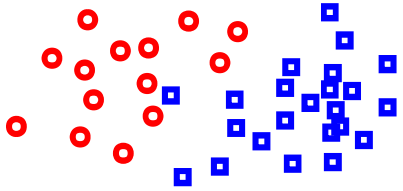
- Gaussian Naïve Bayes



Naïve Bayesian Classification

- Assumption:
  - Follow a normal distribution
- Advantage:
  - Simple and low storage requirements
- Disadvantage:
  - The result can be bad if the group doesn't follow the distribution.

K-Nearest Neighbour

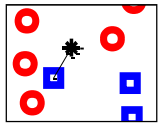


kNN does not build model from the training data.

Consider a two class problem where each sample consists of two measurements (*x*,*y*).

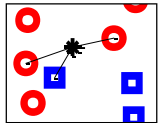
For a given query point q, assign the class of the nearest neighbour.

k = 1



Compute the *k* nearest neighbours and assign the class by majority vote.

k = 3



K-Nearest Neighbour

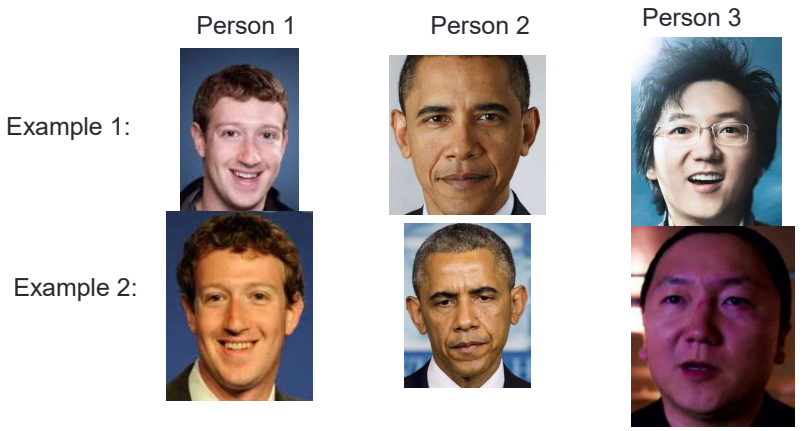
- Expensive
  - To determine the nearest neighbour of a query point *q*, must compute the distance to all *N* training examples
- Storage Requirements
  - Must store all training data
- High Dimensional Data
  - "Curse of Dimensionality"
    - Required amount of training data increases exponentially with dimension
    - Computational cost also increases dramatically

Unsupervised Learning

- Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.
- Goal: group data with similar structures together.

Unsupervised Learning – Grouping Face Data

- Infer the labels



Unsupervised Learning

- K-means algorithm
- Gaussian mixture models
- Hierarchical clustering

K-means clustering

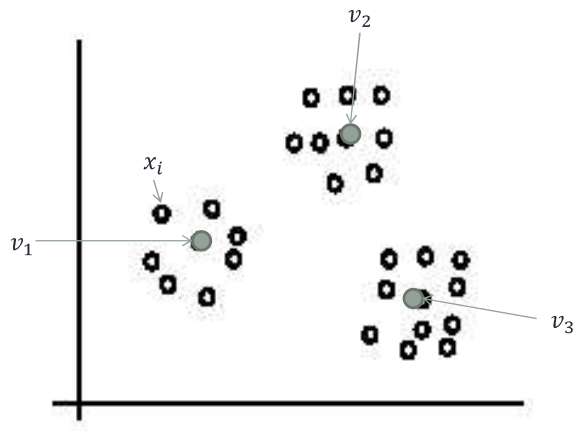
- It partitions data points into K disjoint subsets S_j containing data points so as to minimize the sum-of-squares criterion

$$J = \sum_{k=1}^{K} \sum_{i=1}^{n} I_{ik} ||x_i - v_k||^2$$

- where x_i is a vector representing the i[th] data point, I_{ik} is an indicator function and v_k is the centroid of the kth cluster.

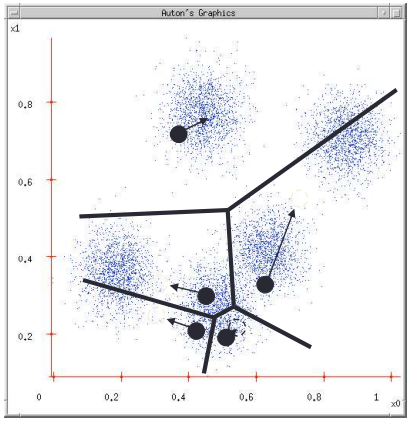
K-means clustering

- 3 natural clusters.



K-means clustering

- K-Means (k , data)
- Randomly choose k cluster center locations (centroids).
- Loop until convergence
  - Assign each point to the cluster of the closest centroid.
  - Re-estimate the cluster centroids based on the data assigned to each.

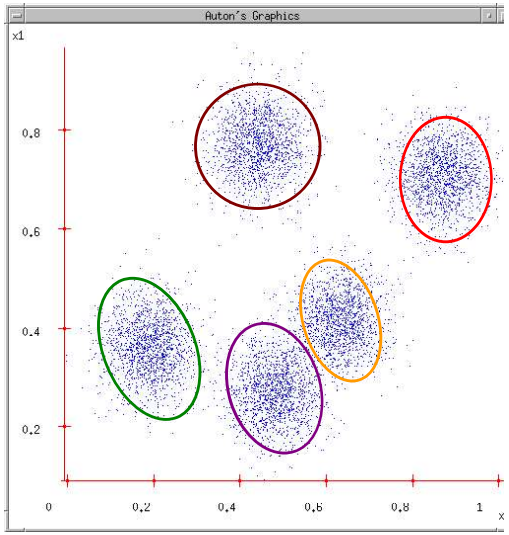


K-means clustering

- ***Disadvantage***
- *Very* sensitive to the initial points.
  - Do many runs of k-Means, each with different initial centroids.
- Must manually choose k.
  - Learn the optimal k for the clustering. (Note that this requires a performance measure.)

Gaussian mixture model

- Assume that data are generated from a mixture of Gaussian distributions
- For each Gaussian distribution
  - Center: μ_i
  - Variance: Σ_i
- For each data point
  - Determine membership



Gaussian mixture model

- The probability given in a mixture of K gaussians is:

$$p(x) = \sum_{j=1}^{K} w_j N(x|\mu_j, \Sigma_j)$$

- where w_j is the prior probability (weight) of the jth Gaussian

$$\sum_{j=1}^{K} w_j = 1 \text{ and } 0 \leq w_j \leq 1$$

K-means v.s. Gaussian mixture model

- Difference between K-means and Gaussian mixture model:
- Membership term:
  - K-means: deterministic
  - Gaussian: stochastic
- Distance function:
  - K-means: without variance
  - Gaussian: with variance

Hierarchical Clustering

- Clustering obtained by cutting the dendrogram at a desired level: each connected component forms a cluster.