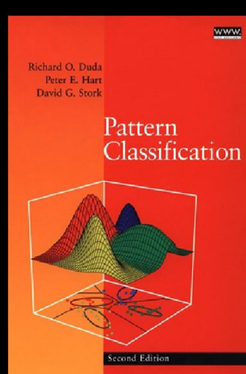
Introduction to Machine Learning
Fall 2013

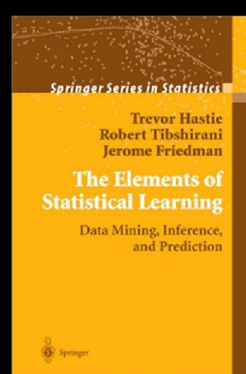
K-means (14)

Koby Crammer
Department of EE
Technion

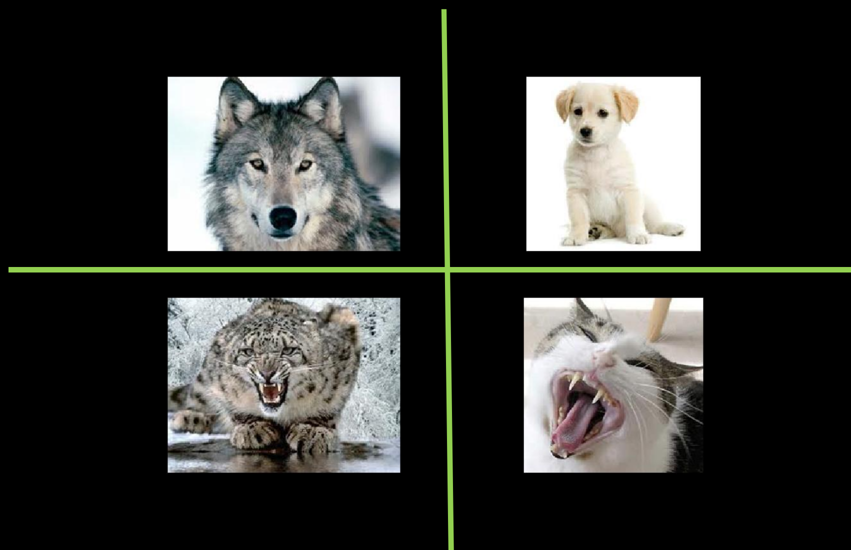
Section 10.4.3



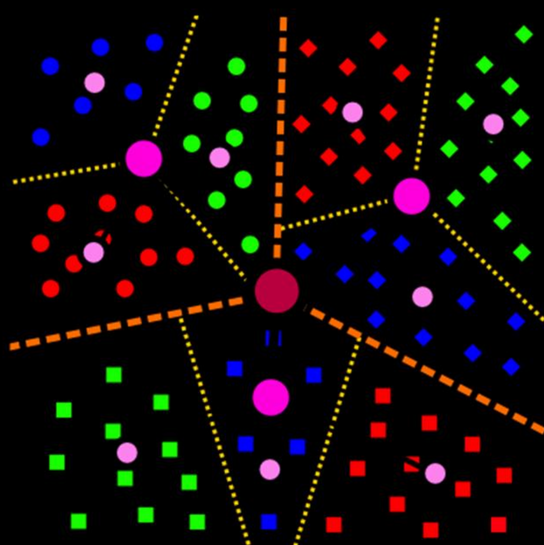
Section 14.3.6



Group Animals

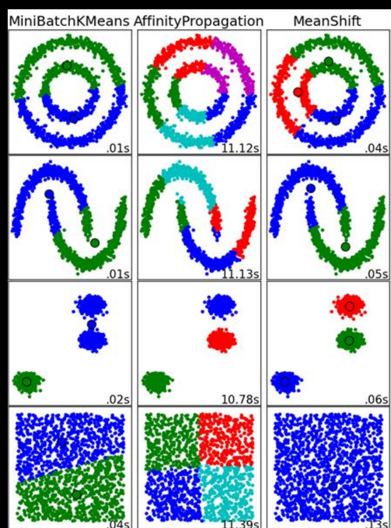


K-Means



4

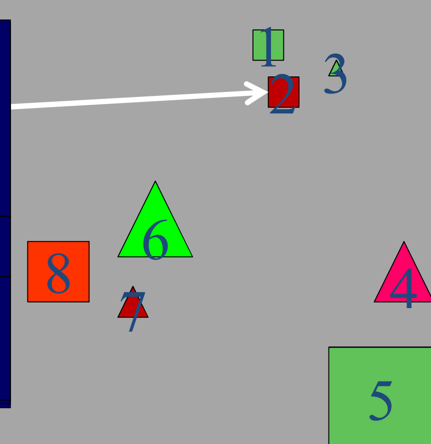
Its all about Assumptions!



5

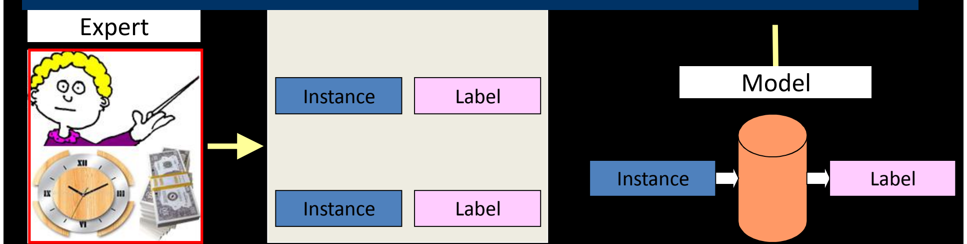
Divide the objects into groups

[redacted]



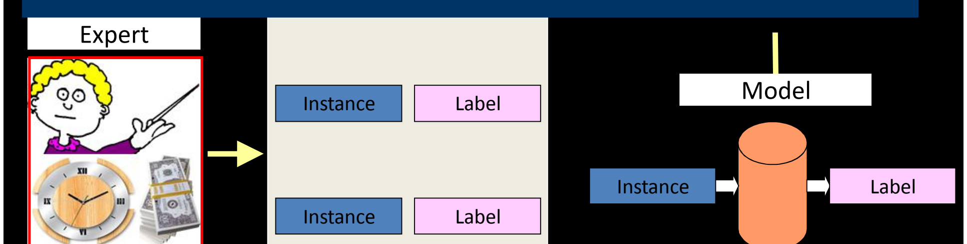
Supervised Learning

- More examples improve performance
- But are more expensive to obtain
- How to choose models?

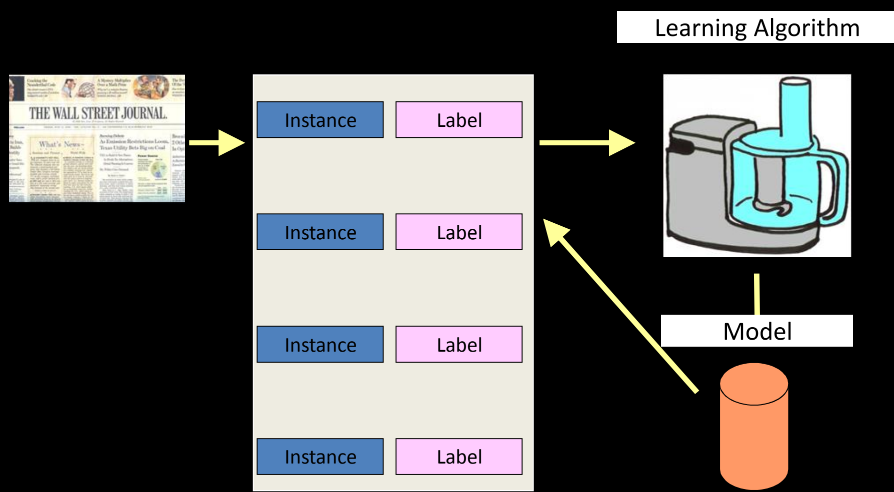


Supervised Learning

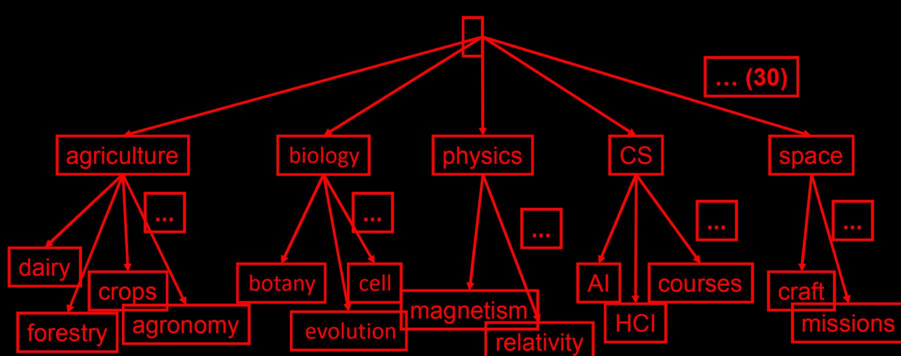
- More examples improve performance
- But are more expensive to obtain
- How to choose models?



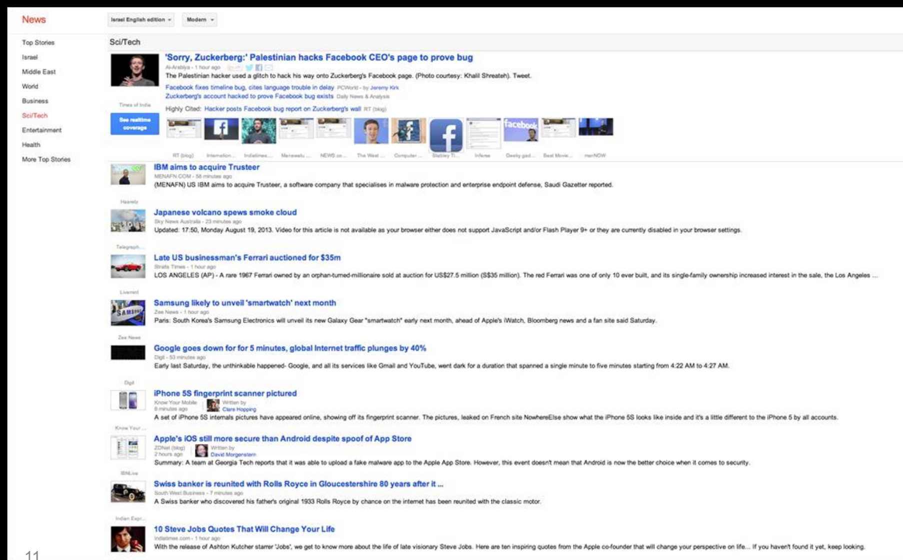
Unsupervised Learning



www.yahoo.com/Science



News Items



11

Topics

- Representation
  - Objects
  - Similarity, distance
- No of Clusters
  - Fixed
  - Too small or too large ?
  - Flat, Dynamical, hierarchical

Soft vs hard clustering


El Al plane returns to airport gate to save sick girl's summer dream

By SAMMY HUDES 08/19/2013 06:45

Like 50 Tweet 50 +1 1 Select Language

The girl, Inbar Chomsky of Rehovot, was one of 30 Israeli children with cancer on her way to Camp Simcha in the US last week.

El Al airplanes sit on the runway Photo: Ronen Zvulun / Reuters

An El Al passenger plane set to take off for New York City returned to the gate at Ben-Gurion Airport to pick up an 11-year-old cancer patient, whose missing passport was found after she was forced off the flight.

- Hard clustering: each object is assigned to a single cluster simple interpretation.
- Soft clustering: obejcts can be assigned to more than a single cluster, maybe with confidence
- Objects have few aspects, such as sports and politics

Clustering vs "examples"

- Problem:
  - Given objects partition them to coherent clusters
- Problem:
  - Given objects find few examples that represent all objects, such that if all objects are associated with their closest representative, we will get coherent clusters
  - (examples may not be subset of inputs)

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

K-Means

- Given N objects partition them into k objects
- Objective:

$$O(\mu, p) = \sum_{i=1}^{N} \sum_{r=1}^{k} p_{i,r} \left\| \mu_r - x_i \right\|^2$$

- Input: matrix of size d x N
- Output :
  - matrix of size d x k (centroids)
  - Matrix of size N x k (association, p)

K-Means

- If matrix of centroids is fixed, what is best association p?

$$\arg\min_{p} \sum_{i=1}^{N} \sum_{r=1}^{k} p_{i,r} \left\| \mu_r - x_i \right\|^2$$

K-Means

- If matrix of centroids is fixed, what is best association p?
- Function decomposes over objects?

$$\arg\min_{p_i} \sum_{r=1}^{k} p_{i,r} \left\| \mu_r - x_i \right\|^2$$

K-Means

- If matrix of centroids is fixed, what is best association p?
- Function decomposes over objects?
- For each take the closest centroid

$$\arg\min_{p_i} \sum_{r=1}^{k} p_{i,r} \left\| \mu_r - x_i \right\|^2$$

$$p_{i,r} = 1 \Leftrightarrow r = \arg\min_j \left\| \mu_j - x_i \right\|^2$$

K-Means

- If the association matrix p is fixed, what is the best matrix ?

$$\arg\min_{\mu} \sum_{i=1}^{N} \sum_{r=1}^{k} p_{i,r} \left\| \mu_r - x_i \right\|^2$$

K-Means

- If the association matrix p is fixed, what is the best matrix ?
- Function decomposes over clusters

$$\arg\min_{\mu_r} \sum_{i=1}^{N} p_{i,r} \left\| \mu_r - x_i \right\|^2$$

K-Means

- If the association matrix p is fixed, what is the best matrix ?
- Function decomposes over clusters
- Compute derivative

$$\mu_r = \frac{\sum_{i=1}^{N} p_{i,r} x_i}{\sum_{i=1}^{N} p_{i,r}}$$

K-Means

- If the association matrix p is fixed, what is the best matrix ?
- Function decomposes over clusters
- Compute derivative
- Given p we know μ

$$\mu_r = \frac{\sum_{i=1}^{N} p_{i,r} x_i}{\sum_{i=1}^{N} p_{i,r}} = \frac{\sum_{i=1}^{N} p_{i,r} x_i}{N_r}$$

K-Means

- Initialize $\mu \in R^{dxk}$
  - Given μ find p

$$p_{i,r} = 1 \Leftrightarrow r = \arg\min_j \left\| \mu_j - x_i \right\|^2$$

  - Given p find μ

$$\mu_r = \left(1 \Big/ \sum_{i=1}^{N} p_{i,r} \right) \sum_{i=1}^{N} p_{i,r} x_i$$

Kernel K-means

- All calculations depend on inner products
- Assume $k(x,y) = \varphi(x) \bullet \varphi(y)$
- Then

$$\left\|\mu_r - \varphi(x_i)\right\|^2 = \left(\mu_r - \varphi(x_i)\right) \bullet \left(\mu_r - \varphi(x_i)\right)$$

$$= \mu_r \bullet \mu_r - 2\mu_r \bullet \varphi(x_i) + \varphi(x_i) \bullet \varphi(x_i)$$

$$= \mu_r \bullet \mu_r - 2\mu_r \bullet \varphi(x_i) + k(x_i, x_i)$$

Kernel K-means

- We get

$$\left\|\mu_r - \varphi(x_i)\right\|^2 = \mu_r \bullet \mu_r - 2\mu_r \bullet \varphi(x_i) + k(x_i, x_i)$$

- Substitute

$$\mu_r = \frac{\sum_{j=1}^{N} p_{j,r} \varphi(x_j)}{N_r}$$

Kernel K-means

$$\left\|\mu_r-\varphi(x_i)\right\|^2 = \frac{\sum_{j=1}^{N} p_{j,r}\varphi(x_j)}{N_r}\bullet\frac{\sum_{j=1}^{N} p_{j,r}\varphi(x_j)}{N_r}$$

$$-2\frac{\sum_{j=1}^{N} p_{j,r}\varphi(x_j)}{N_r}{}_{r}\bullet\varphi(x_i)+k(x_i,x_i)$$

Kernel K-means

$$\left\|\mu_r-\varphi(x_i)\right\|^2 = \frac{1}{N_r^2}\sum_{l,j=1}^{N} p_{j,r}p_{l,r}\varphi(x_j)\bullet\varphi(x_l)$$

$$-2\frac{1}{N_r}\sum_{j=1}^{N} p_{j,r}\varphi(x_j)\bullet\varphi(x_i)+k(x_i,x_i)$$

Kernel K-means

$$\left\|\mu_r - \varphi(x_i)\right\|^2 = \frac{1}{N_r^2}\sum_{l,j=1}^{N} p_{j,r}p_{l,r}K\left(x_j, x_l\right)$$
$$-2\frac{1}{N_r}\sum_{j=1}^{N} p_{j,r}K\left(x_j, x_i\right) + k\left(x_i, x_i\right)$$