Unsupervised Learning

George Konidaris
gdk@cs.duke.edu

DUKE COMPUTER SCIENCE

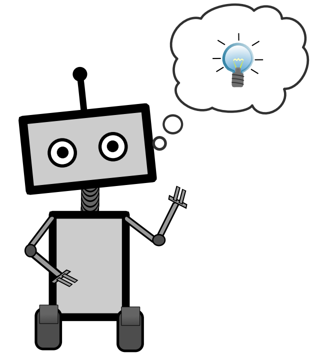
Spring 2015

Machine Learning

Subfield of AI concerned with *learning from data*.

Broadly, using:
- ***Experience***
- To Improve ***Performance***
- On Some ***Task***

(Tom Mitchell, 1997)



Unsupervised Learning

Input:
$X = \{x_1, \ldots, x_n\}$ inputs

Try to understand the *structure of the data.*

E.g., how many types of cars? How can they vary?

Clustering

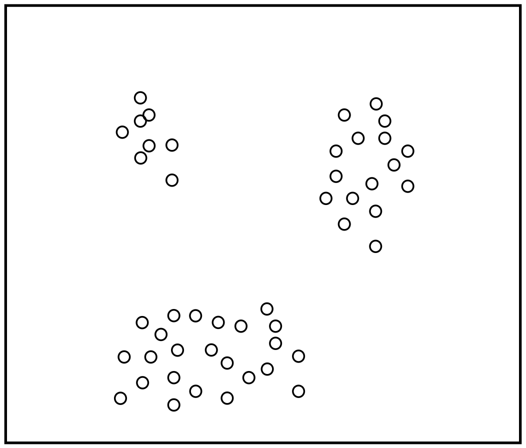
One particular type of unsupervised learning:
- Split the data into discrete clusters.
- Assign new data points to each cluster.
- Clusters can be thought of as *types*.

Formal definition

Given:
- Data points $X = \{x_1, \ldots, x_n\}$,

Find:
- Number of clusters k
- Assignment function $f(x) = \{1, \ldots, k\}$

Clustering

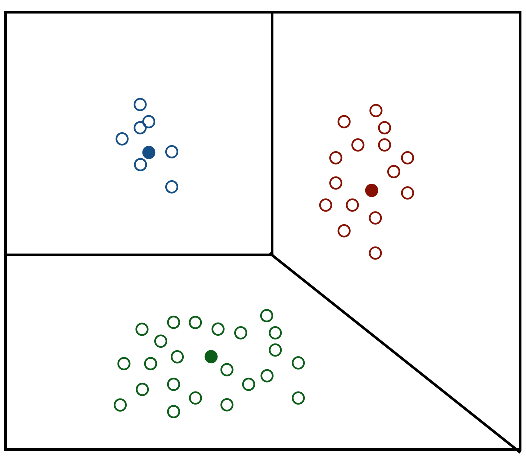


k-Means

One approach:

- Pick *k*
- Place *k* points ("means") in the data
- Assign new point to *i*th cluster if nearest to *i*th "mean".

k-Means



k-Means

Major question:

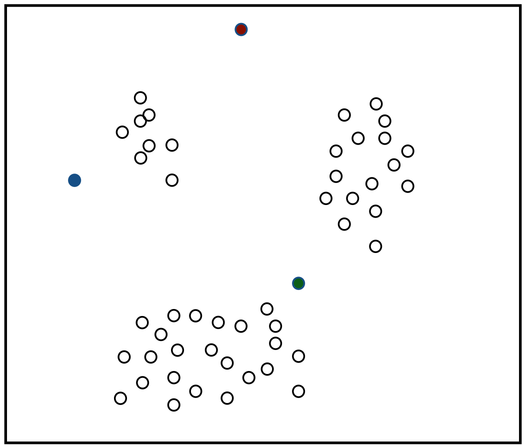
- *Where to put the "means"?*

Very simple algorithm:

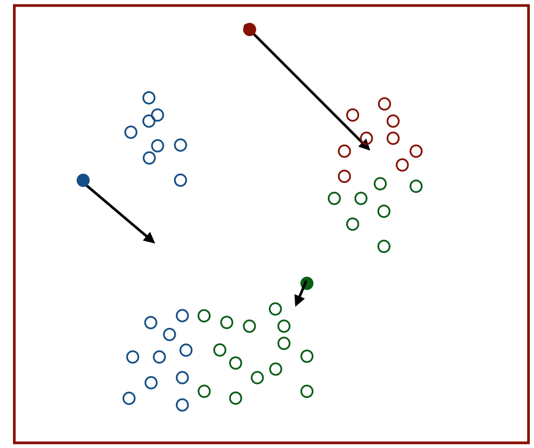
- Place k "means" $\{\mu_1, ..., \mu_k\}$ at random.
- Assign all points in the data to each "mean"
  $f(x_j) = i \text{ such that } d(x_j, \mu_i) \leq d(x_j, \mu_l) \forall l \neq i$
- Move "mean" to mean of assigned data.

$$\mu_i = \sum_{v \in C_i} \frac{x_v}{|C_i|}$$

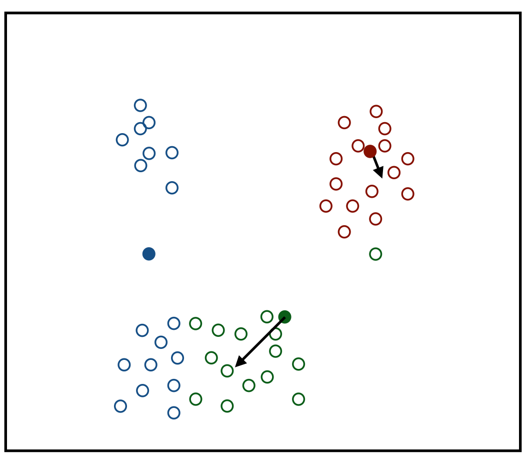
k-Means



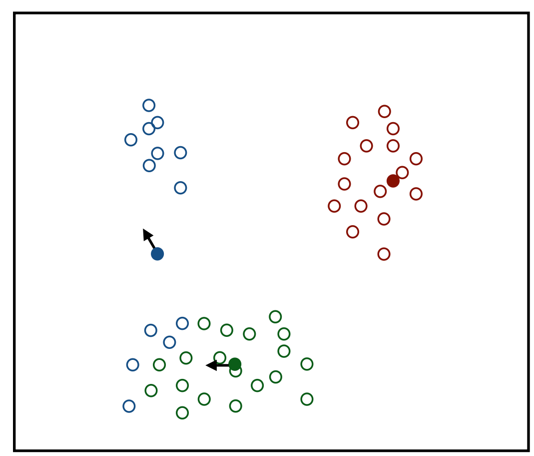
k-Means



k-Means



k-Means



k-Means

Remaining questions …

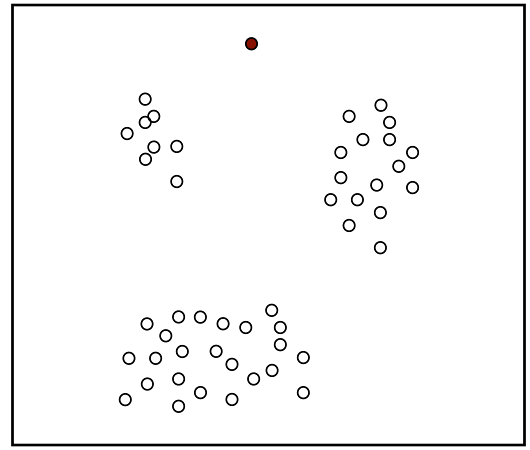
How to choose *k*?

What about bad initializations?

Broadly:
- Use a quality metric.
- Look through *k*.
- Random restart initial position.

Density Estimation

Clustering: can answer *which cluster,* but not *does this belong*?



Density Estimation

Estimate the *distribution the data is drawn from*.

This allows us to evaluate the probability that a new point is drawn from the same distribution as the old data.

Formal definition

Given:
- Data points $X = \{x_1, \ldots, x_n\}$,

Find:
- PDF P(X)

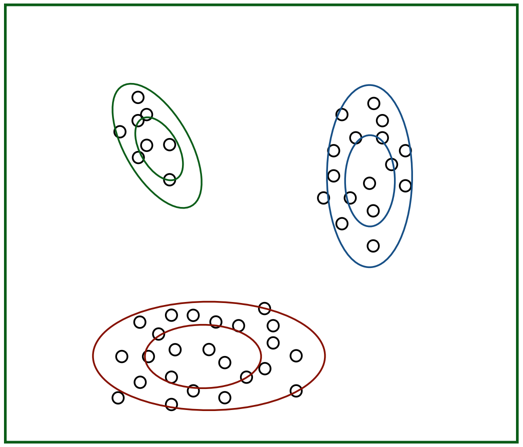
GMM

Simple approach:
- Model the data as a mixture of Gaussians.

Each Gaussian has its own mean and variance.
Each has its own *weight* (sum to 1).

Weighted sum of Gaussians still a PDF.

GMM



GMM

Algorithm - broadly as before:

- Place k "means" $\{\mu_1, \ldots, \mu_k\}$ at random.
- Set variances to be high.

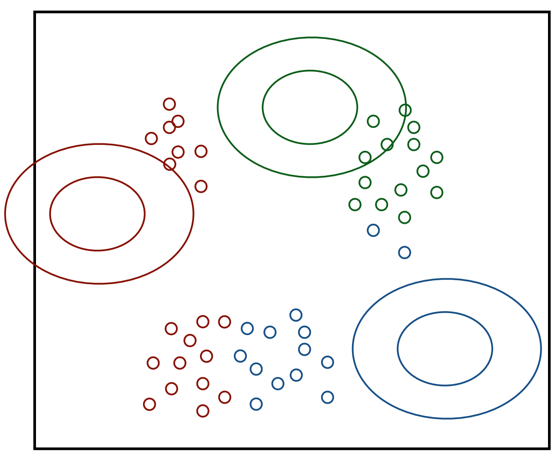
- Assign all points to highest probability distribution.

$$C_i = \{x_v | N(x_v|\mu_i, \sigma_i^2) > N(x_v|\mu_j, \sigma_j^2), \forall j\}$$

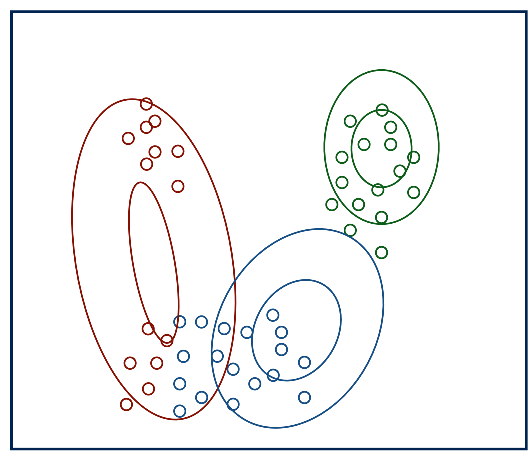
- Set mean, variance to match assigned data.

$$\mu_i = \sum_{v \in C_i} \frac{x_v}{|C_i|} \qquad \sigma_i^2 = \text{variance}(C_i) \qquad w_i = \frac{|C_i|}{\sum_j |C_j|}$$

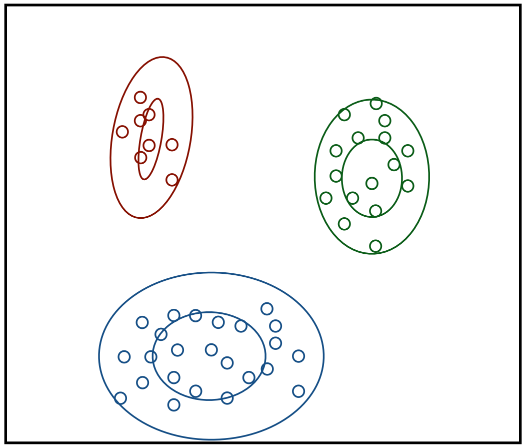
GMM



GMM



GMM



GMM

Major issue:

- How to decide between two GMMs?
- How to choose *k*?

General statistical question: model selection.
Several good answers for this.

Simple example: **Bayesian information criterion (BIC).**
Trades off model complexity (k) with fit (likelihood).

$$-2\log L + k\log n$$

likelihood (L) · # parameters in model (k) · # data points (n)

Application: Novelty Detection

Intrusion detection - when is a user behaving *unusually*?

First proposed by Prof. Dorothy Denning in 1986.
(1995 ACM Fellow)