Clustering

Outline

- Why cluster data?
- Clustering as unsupervised learning
- Clustering algorithms
  - k-means, k-medoids
  - agglomerative clustering
  - Brown's clustering
  - Spectral clustering
- Cluster evaluation measures
  - Purity
  - Normalised Mutual Information
  - Rand Index
  - B-CUBED
  - Precision, Recall, F-score

Why cluster data?

- Data mining has two main objectives:
  - Prediction: classification, regression etc.
  - Description: pattern mining, rule extraction, visualisation, *clustering*
- Clustering is:
  - Unsupervised learning
    - no label data is required (consider classification algorithms we discussed so far in the lectures which are supervised algorithms)

Unsupervised Learning

- Supervised learning
    - labels for training instances are provided
- Unsupervised learning
    - no labels for training instances are provided
- Semi-Supervised learning
    - Both labeled and unlabeled training instances are provided
- What can we learn about training data if we do not have any labels?
    - The similarity and distribution of the features can still be learnt and this can be used to create rich feature spaces for supervised learning (if required)

Clustering: Example

Headlines

More Headlines

Coronavirus: Boris Johnson announces plan for 'delay' phase

Daily Mail · 1 hour ago

- Coronavirus: Boris Johnson to hold emergency Cobra meeting
  BBC South East Wales · 7 hours ago
- BREAKING: UK cases of coronavirus rise to 319
  Sky News · 6 hours ago
- Coronavirus brings a reminder of the iron law of politics
  Financial Times · 4 hours ago · Opinion
- Nigel Farage: Yes, Protecting Us All from an Epidemic Should be Prioritized Over the Economy | Opinion
  Newsweek · 2 hours ago · Opinion

View Full coverage

General Remarks

- A single dataset can be clustered into several ways
- There is no single right or wrong clustering
  - Simply different views of the same data
- how to measure the quality of clustering algorithm?
  - Two ways
    - Compare clusters produced by clustering algorithm against some reference (gold standard) set of clusters (direct evaluation)
    - Use the clusters for some other (eg. supervised learning) task and measure the difference in performance of the second task (indirect evaluation)

Clustering as Optimisation

- Given a dataset $\{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ of N instances represented as d dimensional real vectors ($\mathbf{x}_i \in \mathbb{R}^d$), partition these N instances into k clusters $S_1, \ldots S_k$ such that some objective function $f(S_1, \ldots S_k)$ is minimised.
- Observations
  - k and f are given
  - f can be similarity between the clusters (good to create dissimilar clusters as much as possible), information gain, correlation and various other such goodness measures (heuristics)

Partioning - k-means algorithm

$$\arg\min_{S_1,\ldots,S_k} \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in S_i} ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2$$

We want to minimize the distance between data instances (x_j) and some cluster centres (μ_i)

$$f(S_1,\ldots,S_k) = \sum_{i=1}^{k} \sum_{\boldsymbol{x}_j \in S_i} ||\boldsymbol{x}_j - \boldsymbol{\mu}_i||^2$$

This objective function is called the *within cluster sum of squares* (WCSS) objective

Partioning - cluster centroids

$$\mu_i = \frac{1}{|S_i|} \sum_{\boldsymbol{x}_j \in S_i} \boldsymbol{x}_j$$

Just compute the centroid (mean) of each cluster and that will give you the cluster centers

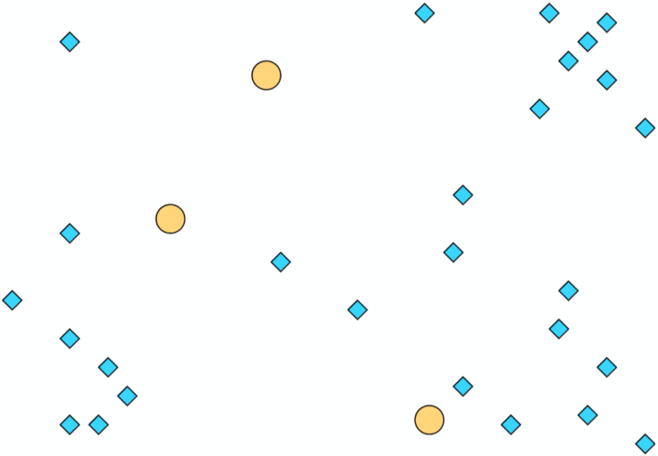
k-Means clustering

- Input
  - The number of clusters k
  - Dataset $\{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ of N instances represented as d dimensional real vectors ($\mathbf{x}_i \in \mathbb{R}^d$)

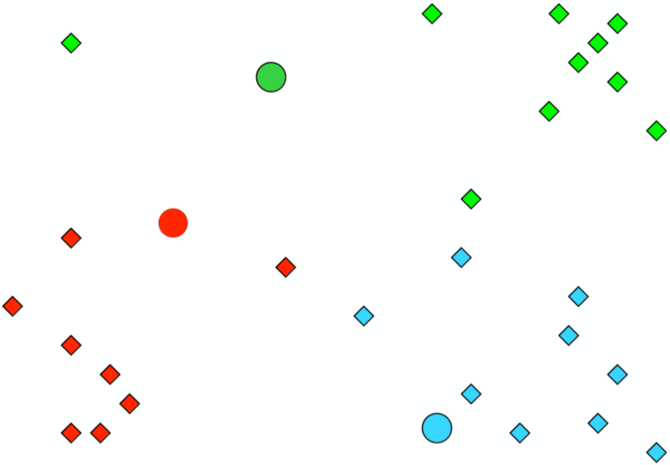
1. Set k instances from the dataset randomly (initial cluster means / centers)
2. Assign all other instances to the closest cluster centre
3. Compute the mean of each cluster
4. Until convergence repeat between steps 2 and 3

convergence = no instances have moved among clusters (often after a fixed number of iterations specified by the user)

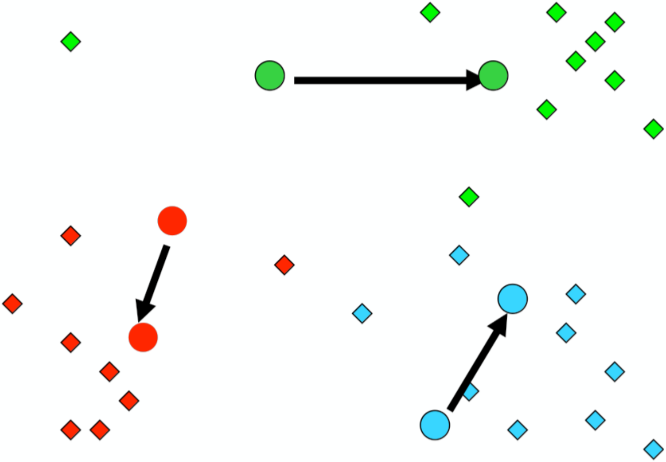
k-means clustering



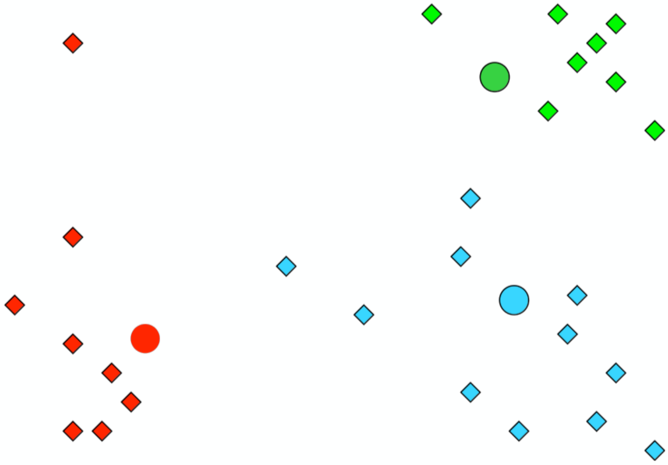
k-means clustering



k-means clustering



k-means clustering



k-means clustering (example)

		c_1 (1,0)	c_2 (1,1)	Assignment
x_1	0,0	$\sqrt{(0-1)^2+(0-0)^2}=\sqrt{1}$	$\sqrt{(0-1)^2+(0-1)^2}=\sqrt{2}$	c_1
x_2	1,0	$\sqrt{(1-1)^2+(0-0)^2}=\sqrt{0}$	$\sqrt{(1-1)^2+(0-1)^2}=\sqrt{1}$	c_1
x_3	1,1	$\sqrt{(1-1)^2+(1-0)^2}=\sqrt{1}$	$\sqrt{(1-1)^2+(1-1)^2}=\sqrt{0}$	c_2
x_4	0,1	$\sqrt{(0-1)^2+(1-0)^2}=\sqrt{2}$	$\sqrt{(0-1)^2+(1-1)^2}=\sqrt{1}$	c_2
x_5	-1,0	$\sqrt{(-1-1)^2+(0-0)^2}=\sqrt{4}$	$\sqrt{(-1-1)^2+(0-1)^2}=\sqrt{5}$	c_1

- $c_1 = \{x_1, x_2, x_5\}; c_2 = \{x_3, x_4\}$
- $c_1 = \{(0,0),(1,0),(-1,0)\}; c_2 = \{(1,1),(0,1)\}$
- $\mu_{c_1} = (0,0); \mu_{c_2} = (0.5, 1)$
- computing clusters using new μ gives the same clusters