SIGNAL & DATA ANALYSIS IN NEUROSCIENCE 2018
CLUSTERING

Ayala Matzner

biu.sigproc@gmail.com

2

Clustering

- A way of grouping together data samples that are ***similar*** in some way - according to some criteria that you pick
- A form of ***unsupervised learning*** – you generally don't have examples demonstrating how the data *should* be grouped together
- How do we define "similarity"?
  - No single answer – it depends on what we want to find or emphasize in the data.
  - The similarity measure is often more important than the clustering algorithm used – don't overlook this choice!

3

The distance function

- Euclidean distance $\text{distance}(x,y) = \sqrt{(\sum_i (x_i - y_i)^2)}$
- Squared Euclidean distance $\text{distance}(x,y) = \sum_i (x_i - y_i)^2$
- City-block (Manhattan) Distance $\text{distance}(x,y) = \sum_i |x_i - y_i|$
- Pearson linear correlation $\text{distance}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$

4

K - means

1) Choose a number of clusters *k*

2) Initialize cluster centers $\mu_1, \ldots \mu_k$
- Could pick *k* data points and set cluster centers to these points
- Or could randomly assign points to clusters and take means of cluster

3) For each data point, compute the cluster center it is closest to (using some distance measure) and assign the data point to this cluster

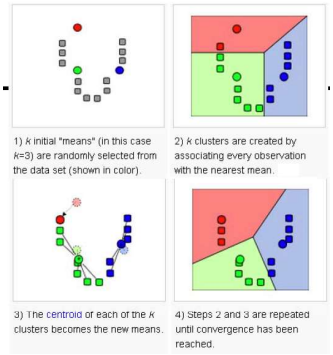
x_i belongs to cluster I if $d(x_i, m_I) < d(x_i, m_j)$, $j \neq I$

4) Re-compute cluster centers (mean of data points in cluster)

5) Stop when there are no new re-assignments

5

K – means example



1) *k* initial "means" (in this case *k*=3) are randomly selected from the data set (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean.

3) The centroid of each of the *k* clusters becomes the new means.

4) Steps 2 and 3 are repeated until convergence has been reached.

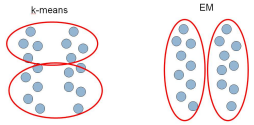
K – means notes

- There are various methods to define distance
- Disadvantages of K-means:
  - Need to specify number of clusters (k) in advance. What if it is unknown?
  - The result may depend on the initial clusters.
  - The algorithm works best on data which contains spherical clusters; clusters with other geometry may not be found.
  - Can't handle outliers and very noisy data
  - Applicable only when mean is defined (e.g. not for categorical data).
- Advantages:
  - Low computing complexity.
  - Easy to implement.
- Matlab: kmeans

7

Model based clustering

- Algorithm optimizes a probabilistic model criterion
- Clustering is usually done by the Expectation Maximization (EM) algorithm
- EM provides soft decision in contrast to the hard decision (p=1 or p=0) k-means – each point belongs with some probability to ALL clusters



Gaussian mixture models (GMM, MOG)

- Generally known as Expectation Maximization (EM): a method for finding ML estimates of parameters in statistical models, where the model depends on unobserved variables.
- Algorithm:

1. Expectation: $p(x_n,g_k) = p(x_n \text{ belongs to Gaussian } g_k)$:

$p(x_n,g_k) = g_k(x_n)/\sum_{i=1,..,K} g_i(x_n)$,

$w(x_n,g_k) = p(x_n,g_k) /\sum_{i=1,..,N} p(x_i,g_k)$

2. Maximization:

Update g_k's center: $m_k = \sum_{i=1,..,N} w(x_i,g_k) \cdot x_i$

Calculate gaussian cov: $V_k = \sum_i w(x_i,g_k) \cdot (x_i - m_k) \cdot (x_i - m_k)'$

3. If m_k or V_k changed, repeat expectation step.
Else end.

Example: exam 2007

One thousand elves, dwarves, ogres and goblins are assessed using 50 "almost" normal parameters (heights, ear shape, weight, decay of teeth, etc.). What should be done to identify one "typical" member of each of the species?

a. K-Means followed by PCA.
b. PCA followed by K-Means.
c. Mutual information followed by maximum likelihood estimator (MLE).
d. Maximum likelihood estimator (MLE) followed by mutual information.