Unsupervised Learning

Clustering

Ioannis Patras

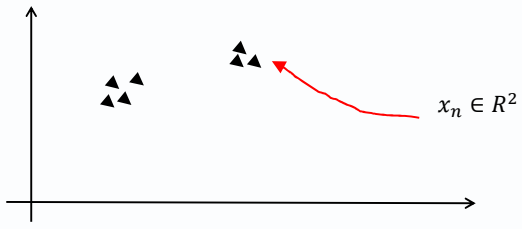
1

Unsupervised Learning

By contrast to supervised learning the training set consists of only data points (no 'target data').

Goal: Find some structure in the data

$$X = \{x_n | n = 1 \dots m\}$$



2

Clustering

Given the hue and size (features) of fruits, group them into clusters

Possible uses: Find hidden structure in the data

- news clustering, genes clustering, image clustering, audio clustering.

Slide no: 0-3

3

K - means clustering

Given: $X = \{x_n | i = 1 \dots m\}$ $x_n \in R^N$ and the number of clusters K

Find: $\{c_n | n = 1 \dots m\}, c_n \in [1 \dots K]$ and $\{\mu_k | k = 1 \dots \mathrm{K}\}, \mu_k \in R^N$

The K-Means clustering algorithm works as follows. First initialize K centres $\boldsymbol{\mu}_k$ (for example by picking K input samples randomly) then iteratively repeat the following steps.

$\mathrm{Step} - 1$: classify x_n to the cluster k with the nearest centre μ_k

$c_n = k$ (or $x_n \in C_k$) if $\|x_n - \mu_{c_n}\| < \|x_n - \mu_j\|$ for all $j \neq k$

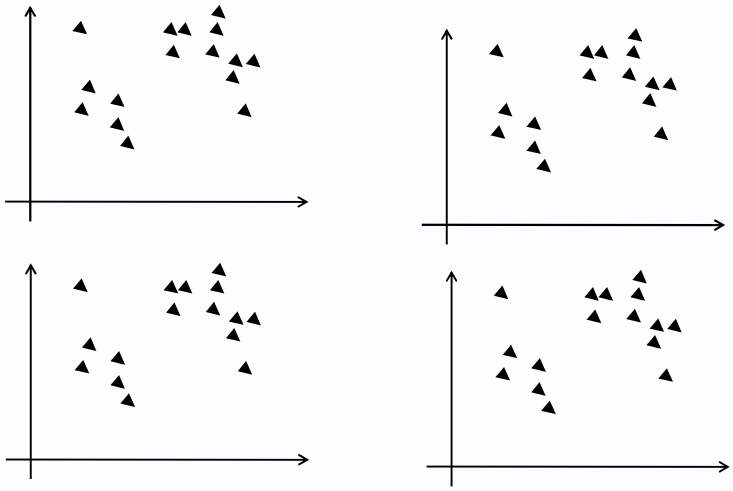
$\mathrm{Step} - 2$: re$-$compute μ_k to be the mean of the class C_k

$$\mu_k = \frac{\sum_{x_n \in C_k} x_n}{\sum_{x_n \in C_k} 1} = \frac{1}{|C_k|} \sum_{x_n \in C_k} x_n$$

Until the solution stabilizes (i.e. until clusters of x_n stop changing in Step 1).

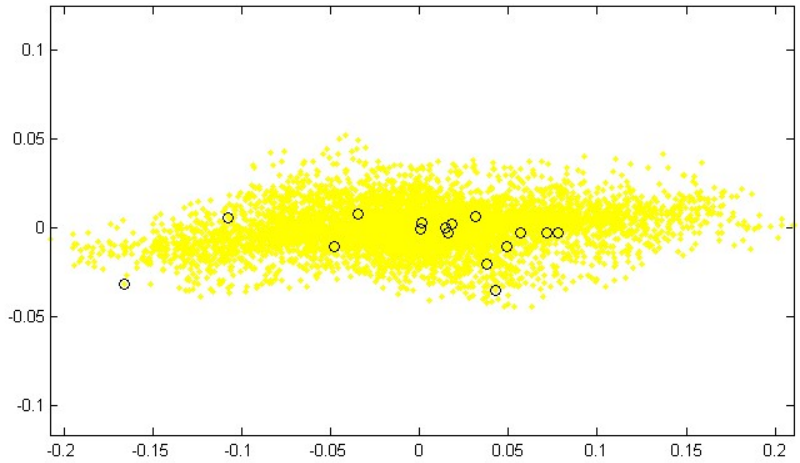
4

K - means clustering (example)



5

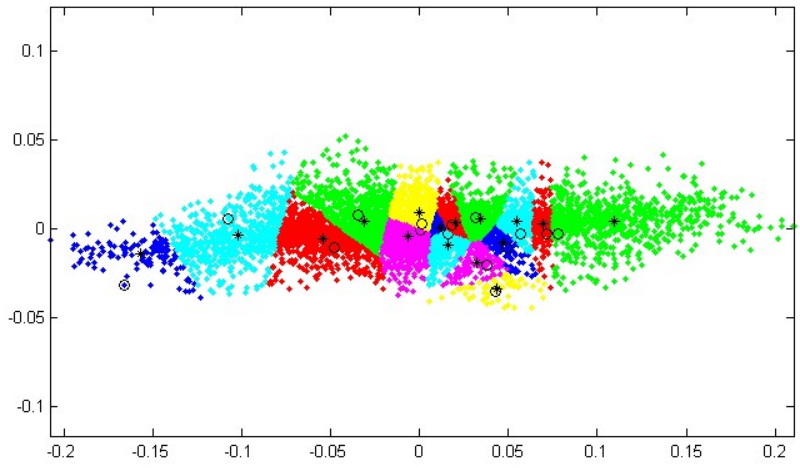
K - means: Initialisation



We have used random initialisation – more sophisticated schemes exist

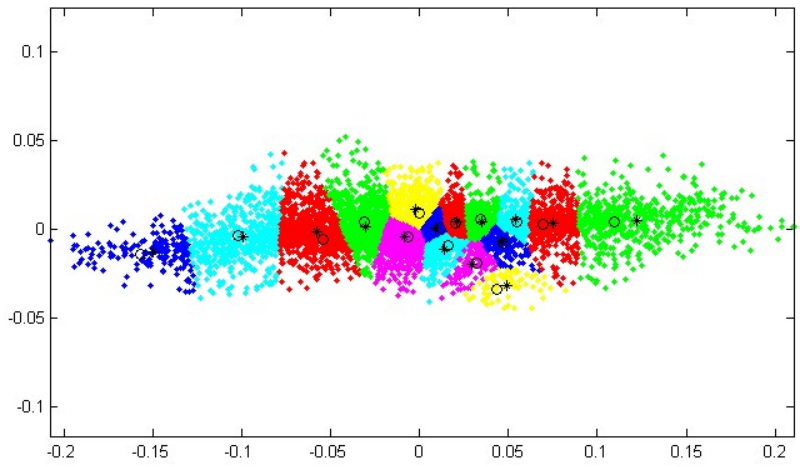
6

K - means: 1st iteration



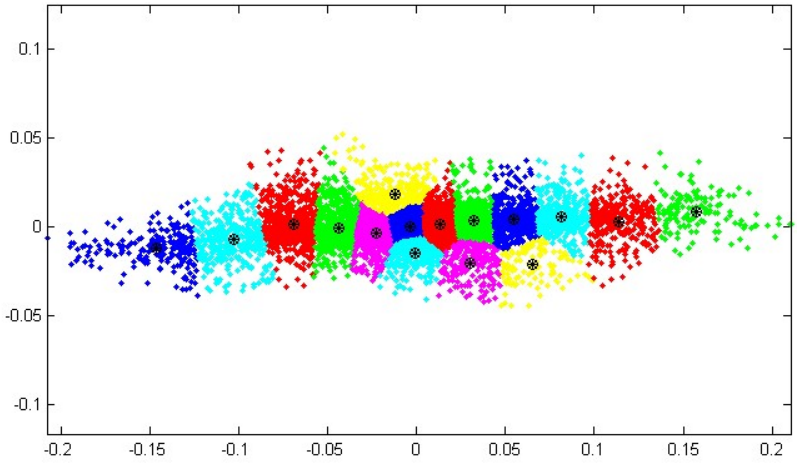
7

K - means: 2nd iteration



8

K - means: converged (20th iteration)



9

Convergence of K - means

We can now show that the two steps of K-means optimise a cost function

$$J(c_1, \ldots, c_n, \mu_1, \ldots, \mu_K) = \sum_n \|x_n - \mu_{c_n}\|^2$$

Where $c_n \in [1 \ldots K]$ indicates the cluster to which x_n belongs.

Non convex function, may have local minima, difficult to optimize.

K-means follows an optimisation scheme called co-ordinate descent (or ascent if the objective needs to be maximised)

10

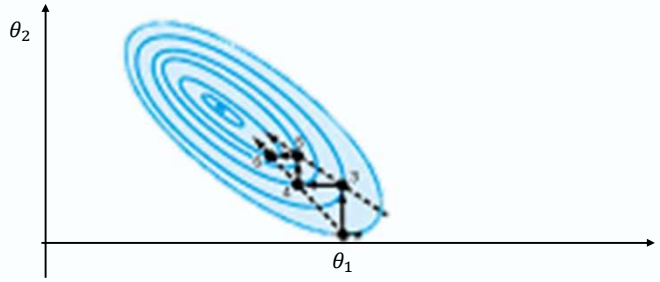
Coordinate descent

Minimize $J(\theta_1, \theta_2)$ by iterating between two steps:

Step 1: Keep θ_2 fixed and optimise w.r.t. θ_1, i.e., $\theta_1^{k+1} = argmin\, J(\theta_1, \theta_2^k)$

Step 2: Keep θ_1 fixed and optimise w.r.t θ_2, i.e., $\theta_2^{k+1} = argmin\, J(\theta_1^{k+1}, \theta_2)$

Until convergence



Given some conditions, the procedure converges to a local minimum

11

"Optimality" of K - means (1)

Cost function $J(c_1, \ldots, c_m, \mu_1, \ldots, \mu_K) = \sum_n \|x_n - \mu_{c_n}\|^2$

Step 1: Optimise w.r.t $c_1, \ldots, c_m$

$$c_n^* = \text{argmin}_{c_n} \sum_{n=1} \|x_n - \mu_{c_n}\|^2 = \text{argmin}_{c_n} \|x_n - \mu_{c_n}\|^2$$

Thus to minimise the cost with respect to classifications we choose $\mathbf{c_n}$ to be the cluster k for which $||x_n - \mu_k||^2$ is smallest (i.e. we choose the closest centre).

12

"Optimality" of K - means (1)

Cost function $J(c_1, \dots, c_m, \mu_1, \dots, m_k) = \sum_{n=1}^{m} \left|\left|x_n - \mu_{c_n}\right|\right|^2$

Step 2: Optimise w.r.t μ_k by setting the derivative wrt it to zero.

$$\nabla_{\mu_k} J(\dots) = \sum_{x_n \in C_k} -2\mu_k(x_n - \mu_k)$$

$$\nabla_{\mu_k} J(\dots) = 0 \Rightarrow \mu_k = \frac{\sum_{x_n \in C_k} x_n}{\sum_{x_n \in C_k} 1} = \frac{1}{|C_k|} \sum_{x_n \in C_k} x_n$$

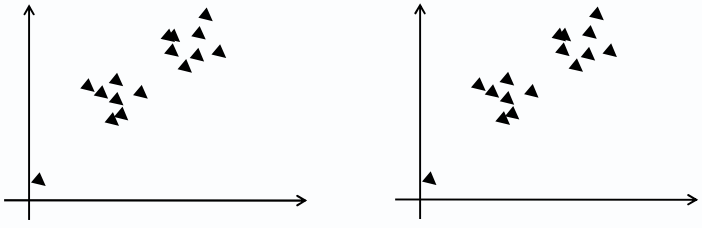
Where $|C_k|$ are the number of points assigned to cluster k

That is, the μ_k that minimizes the cost function is the average of the points that belong to cluster k.

13

Local minima

K-means converges only to a local minimum.



Usually K-means is run several times with different initialisations.

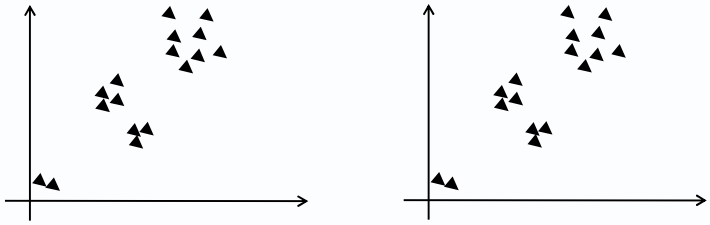
The quality of each solution is determined by evaluating the cost function

The solution leading to the lowest cost is chosen.

14

How many clusters?

Ambiguous! Not well posed problem.

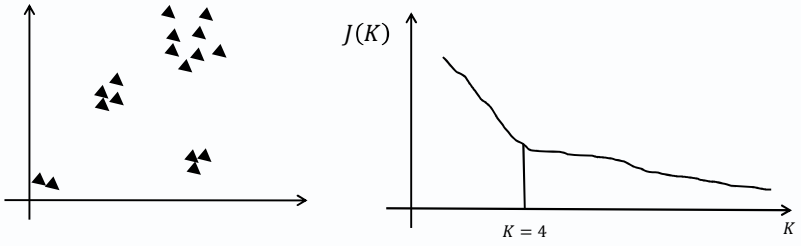


Cost function is monotonic wrt K. Larger K leads to smaller J

15

How many clusters?

1. Plot cost as function of K. Cost function is monotonic wrt K. Larger K leads to smaller J. In some cases, the shape may give a hint for the "correct" value of K.



16

How many clusters?

2. Evaluate on the application. Obtain different clusters for different values of K. Calculate a measure $J'(K)$ that depends on the application

   Example a: News clustering:

   Evaluate: Present to actual users clustering results with different number of clusters. Asses whether they prefer fewer clusters (i.e. Broader topics) or more clusters (i.e. More specialized topics)

   Example b: A recommendation system cluster users into groups and gives recommendations according to which cluster it think that a user belongs.

17

How many clusters?

3. Calculate a measure that penalises complex solutions / many clusters

$$J'(K) = \text{dK lnm+ } \sum_n x_n \ - \mu_{c_n}|_2^2$$

where d is the dimensionality

K the number of clusters

m the number of data points.

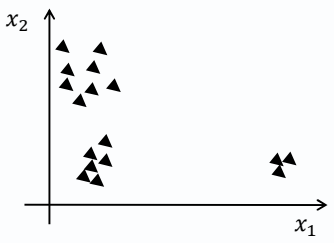
Related to Bayesian Information Criterion (BIC).

18

K-means issues: Data scaling

Is sensitive to data scaling.

$x_1 \leftarrow \alpha x_1$ can move the cluster of the points in the right arbitrarily to the left or right.



Normalisation (stretching) $x_1 \leftarrow \frac{x_1 - min_n\left\{x_1^{(n)}\right\}}{\max_n\left\{x_1^{(n)}\right\} - min_n\left\{x_1^{(n)}\right\}}$

Normalisation (by variance) $x_1 \leftarrow \sigma^{-1} x_1$

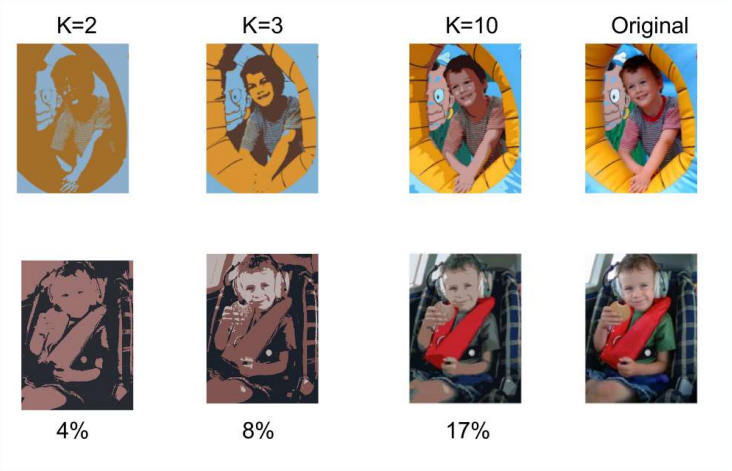
19

Applications: Image Segmentation

Group together pixels that are homogenous in their properties (e.g., colour)

20

Applications: Image Segmentation



Represent each pixel with the mean of its cluster

$x \in R^3$ 3 colour channels

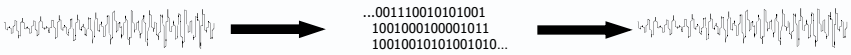
21

Applications: K - means for VQ

Problem: *What is the best (under MMSE) way to quantize vectors of* $x \in R^d$ *using N bits per sample?*

N bits means 2^N different representatives. (K = 2^N ☺

This is called Vector Quantization (VQ). It was the basis for speech codecs.



Vector Quantization for continuous speech signal:

1. Divide the signal into non overlapping segments of length d.
2. Cluster the segments (each one $\in R^d$) in to K clusters by K-means
3. Represent each segment by the index of the corresponding cluster using N bits coding. At the decoder side, decode using the corresponding cluster mean.

22

Applications: VQ

FIGURE 14.9. *Sir Ronald A. Fisher* $(1890 - 1962)$ *was one of the founders of modern day statistics, to whom we owe maximum-likelihood, sufficiency, and many other fundamental concepts. The image on the left is a* 1024×1024 *grayscale image at* 8 *bits per pixel. The center image is the result of* 2×2 *block VQ, using* 200 *code vectors, with a compression rate of* 1.9 *bits/pixel. The right image uses only four code vectors, with a compression rate of* 0.50 *bits/pixel*

[Figure from Hastie *et al.* book]

23

Summary

- Introduced the simple K-means clustering algorithm
- Limitations
- Convergence, optimality, scaling, number of clusters
- Applications

24