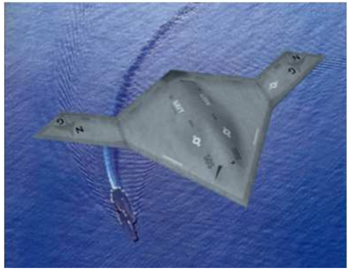
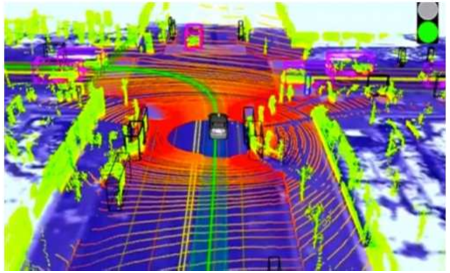
Intelligent Image and Graphics Processing
智能图像图形处理

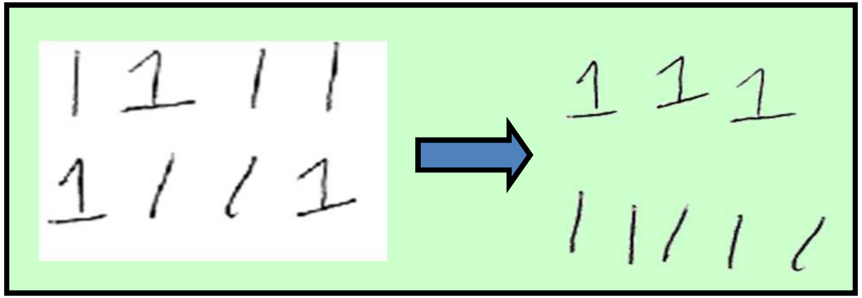
布树辉 bushuhui@nwpu.edu.cn
http://www.adv-ci.com

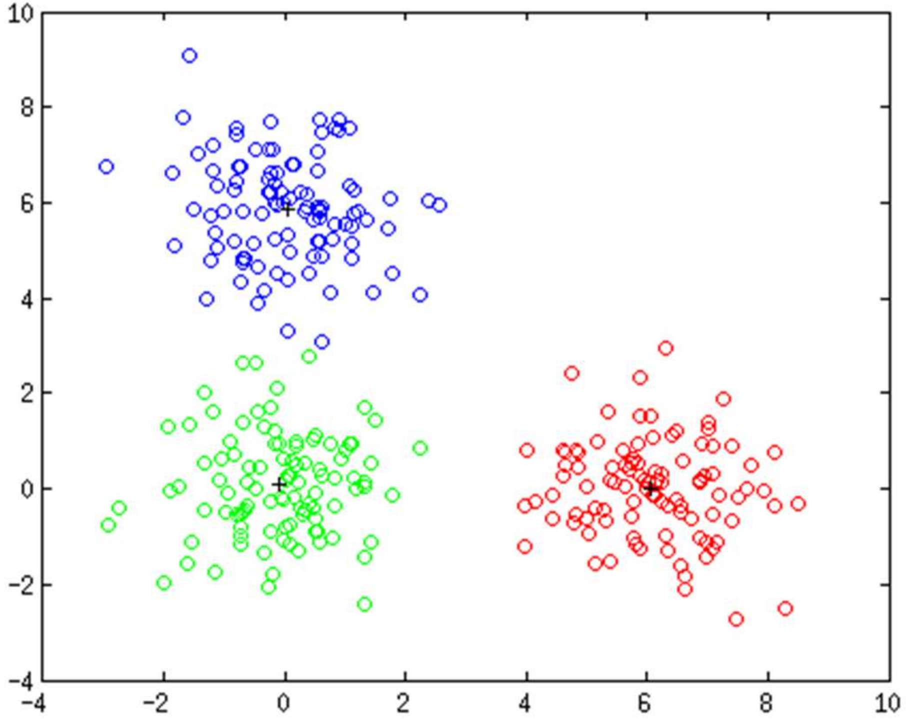

Clustering

Clustering

- Attach label to each observation or data points in a set
- You can say this "unsupervised classification"
- Clustering is alternatively called as "grouping"
- Intuitively, if you would want to assign same label to a data points that are "close" to each other
- Thus, clustering algorithms rely on a **distance metric** between data points
- Sometimes, it is said that the for clustering, the distance metric is more important than the clustering algorithm



Clustering



Distances: Quantitative Variables

Some examples

Identity (absolute) error

$$d_j(x_{ij}, x_{i'j}) = I(x_{ij} \neq x_{i'j})$$

Squared distance

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$$

L_q norms

$$L_{qii'} = [\sum_j |x_{ij} - x_{i'j}|^q]^{1/q}$$

Canberra distance

$$d_{ii'} = \sum_j \frac{|x_{ij} - x_{i'j}|}{|x_{ij} + x_{i'j}|}$$

Correlation

$$\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}$$

Data point:

$$x_i = [x_{i1} \ldots x_{ip}]^T$$

Partitioning Clustering Approach

- Partitioning Clustering Approach
  - a typical clustering analysis approach via iteratively partitioning training data set to learn a partition of the given data space
  - learning a partition on a data set to produce several non-empty clusters (usually, the number of clusters given in advance)
  - in principle, optimal partition achieved via minimizing the sum of squared distance to its "representative object" in each cluster

$$E = \sum_{k=1}^{K} \sum_{\mathbf{x} \in C_k} d^2(\mathbf{x}, \mathbf{m}_k)$$

e.g., Euclidean distance $d^2(\mathbf{x}, \mathbf{m}_k) = \sum_{n=1}^{N} (x_n - m_{kn})^2$

Partitioning Clustering Approach

- Given a *K*, find a partition of *K clusters* to optimize the chosen partitioning criterion (cost function)
  - global optimum: exhaustively search all partitions
- The *K-means* algorithm: a heuristic method
  - K-means algorithm (MacQueen'67): each cluster is represented by the center of the cluster and the algorithm converges to stable centriods of clusters.
  - K-means algorithm is the simplest partitioning method for clustering analysis and widely used in data mining applications.

Partitioning Clustering Approach

Given the cluster number *K*, the *K-means* algorithm is carried out in three steps after initialization:

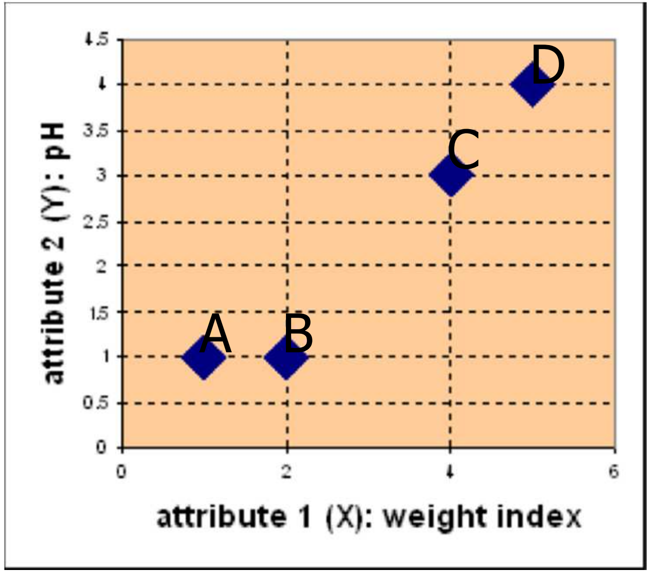
Initialisation: set seed points (randomly)

1) Assign each object to the cluster of the nearest seed point measured with a specific distance metric
2) Compute new seed points as the centroids of the clusters of the current partition (the centroid is the centre, i.e., *mean point*, of the cluster)
3) Go back to Step 1), stop when no more new assignment (i.e., membership in each cluster no longer changes)

Example

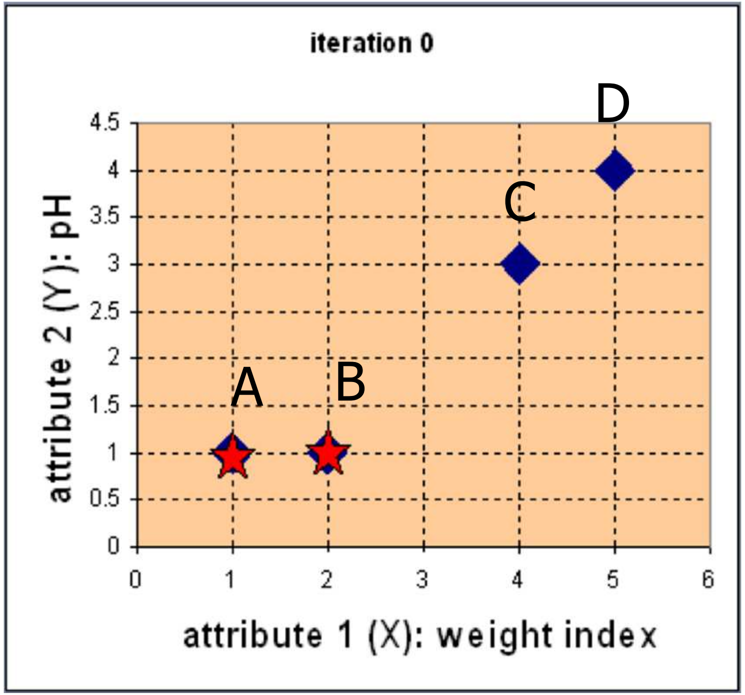
Suppose we have 4 types of medicines and each has two attributes (pH and weight index). Our goal is to group these objects into *K=2* group of medicine.

Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4



Example

- Step 1: Use initial seed points for partitioning



$$c_1 = A, c_2 = B$$

$$\mathbf{D}^0 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 1 & 0 & 2.83 & 4.24 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (2,1) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Euclidean distance

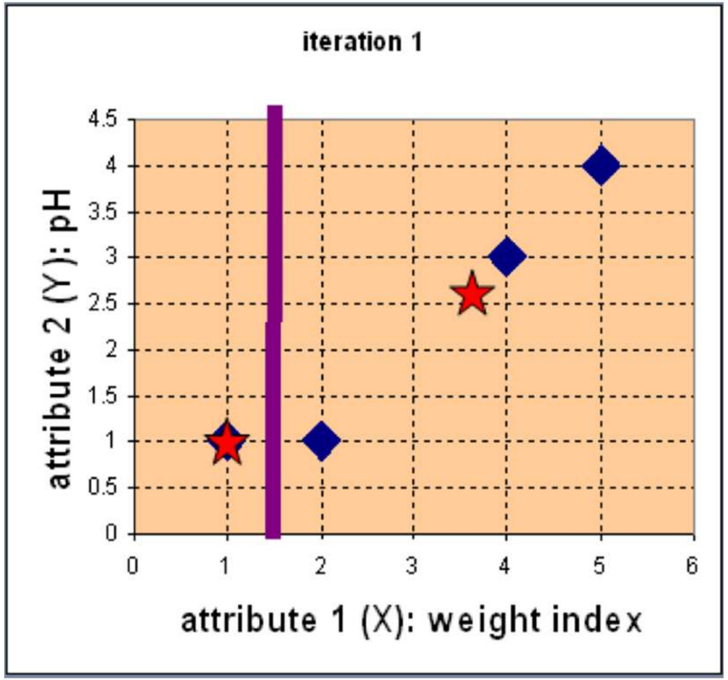
$$d(D, c_1) = \sqrt{(5-1)^2 + (4-1)^2} = 5$$

$$d(D, c_2) = \sqrt{(5-2)^2 + (4-1)^2} = 4.24$$

Assign each object to the cluster with the nearest seed point

Example

- Step 2: Compute new centroids of the current partition



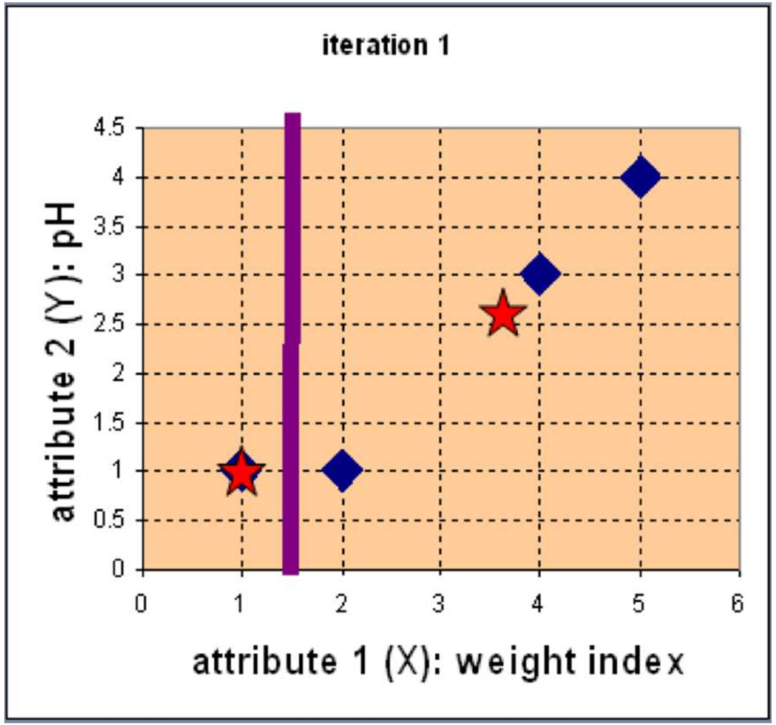
Knowing the members of each cluster, now we compute the new centroid of each group based on these new memberships.

$$c_1 = (1,\ 1)$$

$$c_2 = \left(\frac{2+4+5}{3},\ \frac{1+3+4}{3}\right)$$
$$= \left(\frac{11}{3},\ \frac{8}{3}\right)$$

Example

- Step 2: Renew membership based on new centroids



Compute the distance of all objects to the new centroids

$$\mathbf{D}^1 = \begin{bmatrix} 0 & 1 & 3.61 & 5 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1,1) & group-1 \\ \mathbf{c}_2 = (\frac{11}{3}, \frac{8}{3}) & group-2 \end{matrix}$$

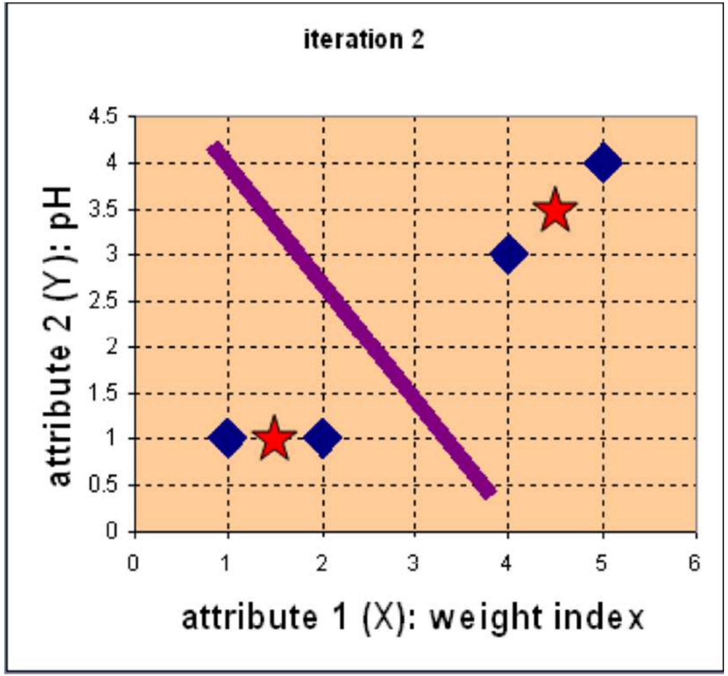
$$\begin{matrix} A & B & C & D & \\ \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

Assign the membership to objects

Example

- Step 3: Repeat the first two steps until its convergence



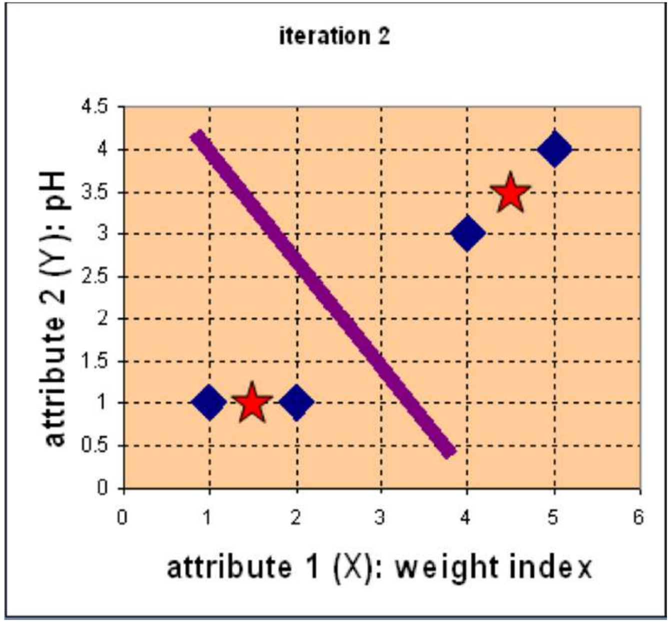
Knowing the members of each cluster, now we compute the new centroid of each group based on these new memberships.

$$c_1 = \left(\frac{1+2}{2}, \ \frac{1+1}{2} \right) = (1\frac{1}{2}, \ 1)$$

$$c_2 = \left(\frac{4+5}{2}, \ \frac{3+4}{2} \right) = (4\frac{1}{2}, \ 3\frac{1}{2})$$

Example

- Step 3: Repeat the first two steps until its convergence



Compute the distance of all objects to the new centroids

$$\mathbf{D}^2 = \begin{bmatrix} 0.5 & 0.5 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix} \quad \begin{matrix} \mathbf{c}_1 = (1\frac{1}{2}, 1) & group-1 \\ \mathbf{c}_2 = (4\frac{1}{2}, 3\frac{1}{2}) & group-2 \end{matrix}$$

$$\begin{matrix} A & B & C & D \end{matrix}$$

$$\begin{bmatrix} 1 & 2 & 4 & 5 \\ 1 & 1 & 3 & 4 \end{bmatrix} \begin{matrix} X \\ Y \end{matrix}$$

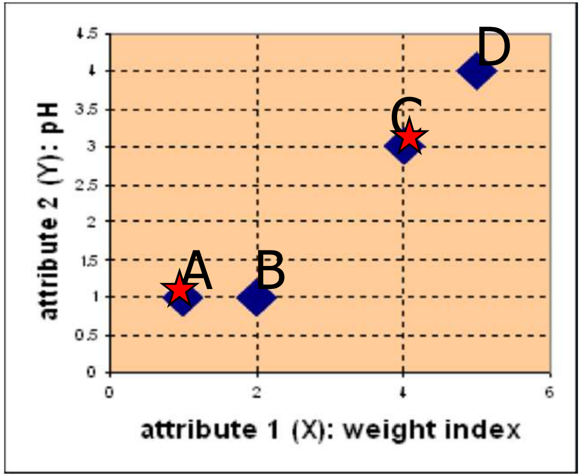
Stop due to no new assignment
Membership in each cluster no longer change

Exercise

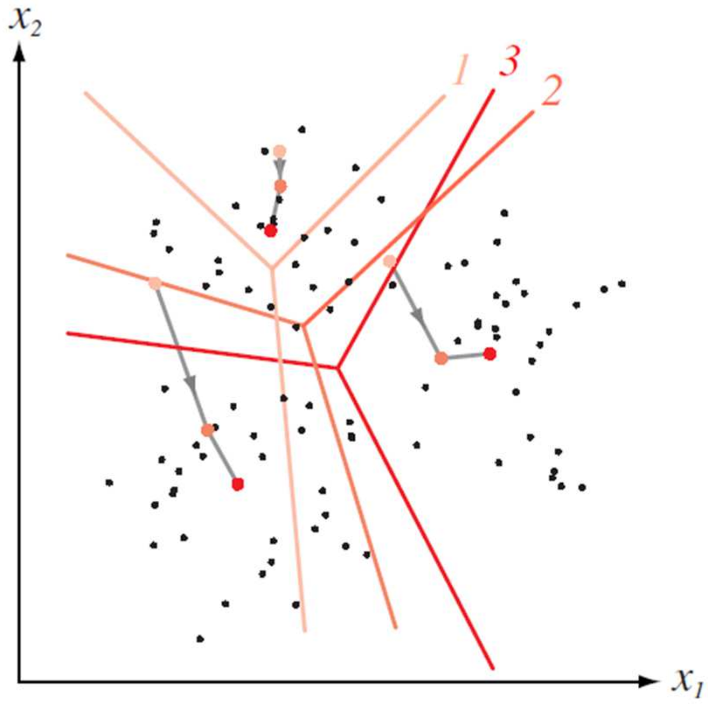
For the medicine data set, use K-means with the Manhattan distance metric for clustering analysis by setting K=2 and initializing seeds as C_1 = A and C_2 = C. Answer three questions as follows:

1. How many steps are required for convergence?
2. What are memberships of two clusters after convergence?
3. What are centroids of two clusters after convergence?

Medicine	Weight	pH-Index
A	1	1
B	2	1
C	4	3
D	5	4



How K-means partitions?



When K centroids are set/fixed, they partition the whole data space into K mutually exclusive subspaces to form a partition.

A partition amounts to a

Voronoi Diagram

Changing positions of centroids leads to a new partitioning.

Application - Colour-Based Image Segmentation Using *K*-means

H&E image

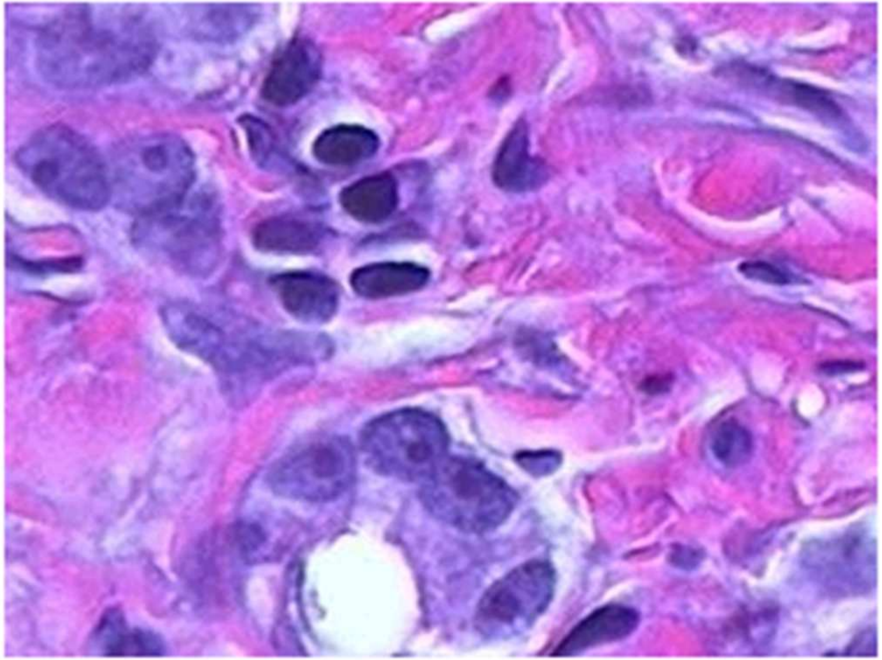


Image courtesy of Alan Partin, Johns Hopkins University

Step 1: Loading a colour image of tissue stained with hemotoxylin and eosin (H&E)

Application - Colour-Based Image Segmentation Using *K*-means

Step 2: Convert the image from RGB colour space to L*a*b* colour space

- Unlike the RGB colour model, L*a*b* colour is designed to approximate human vision.
- There is a complicated transformation between RGB and L*a*b*.

$$(L^*, a^*, b^*) = T(R, G, B).$$

$$(R, G, B) = T'(L^*, a^*, b^*).$$

Application - Colour-Based Image Segmentation Using *K*-means

Step 3: Undertake clustering analysis in the (a*, b*) colour space with the *K*-means algorithm

- In the L*a*b* colour space, each pixel has a properties or feature vector: (L*, a*, b*).
- Like feature selection, L* feature is discarded. As a result, each pixel has a feature vector (a*, b*).
- Applying the *K*-means algorithm to the image in the a*b* feature space where $K = 3$ (by applying the domain knowledge.

Application - Colour-Based Image Segmentation Using *K*-means

Step 4: Label every pixel in the image using the results from *K*-means Clustering (indicated by three different grey levels)

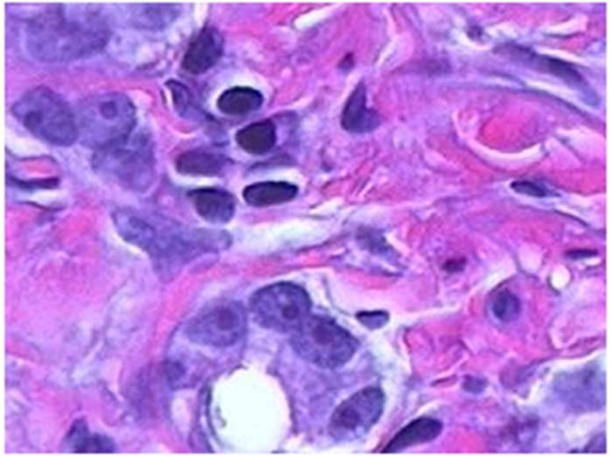
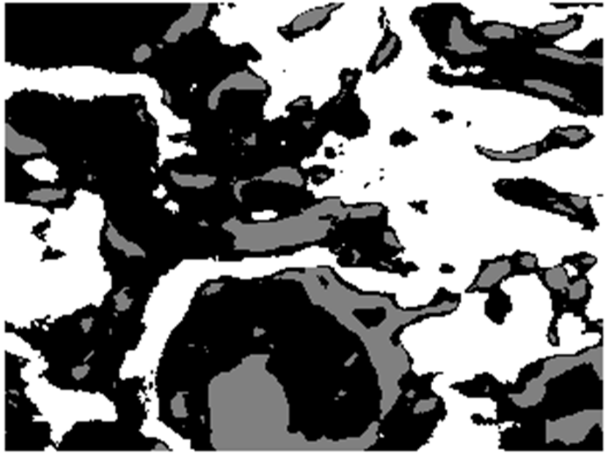


Image courtesy of Alan Partin, Johns Hopkins University



Application - Colour-Based Image Segmentation Using *K*-means

Step 4: Label every pixel in the image using the results from *K*-means Clustering (indicated by three different grey levels)

H&E image

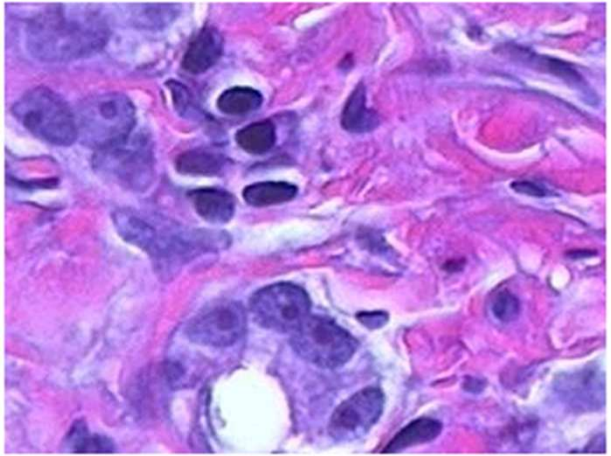
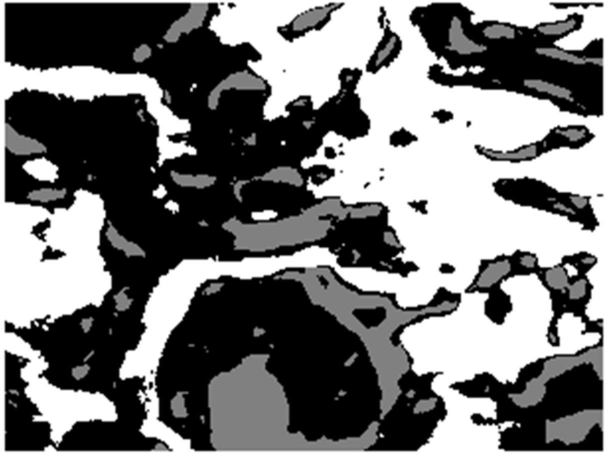


Image courtesy of Alan Partin, Johns Hopkins University

image labeled by cluster index



Application: Learning Feature Representations

1. Normalize inputs:

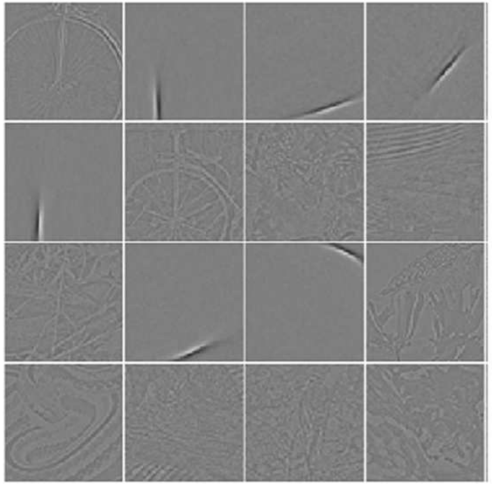
$$x^{(i)} := \frac{x^{(i)} - \text{mean}(x^{(i)})}{\sqrt{\text{var}(x^{(i)}) + \epsilon_{\text{norm}}}}, \forall i$$

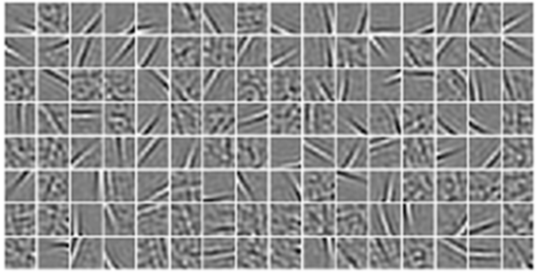
2. Whiten inputs:

$$[V, D] := \text{eig}(\text{cov}(x)); \quad // \text{ So } VDV^\top = \text{cov}(x)$$
$$x^{(i)} := V(D + \epsilon_{\text{zca}} I)^{-1/2} V^\top x^{(i)}, \forall i$$

3. Loop until convergence (typically 10 iterations is enough):

$$s_j^{(i)} := \begin{cases} \mathcal{D}^{(j)\top} x^{(i)} & \text{if } j == \arg\max_l |\mathcal{D}^{(l)\top} x^{(i)}| \\ 0 & \text{otherwise.} \end{cases} \quad \forall j, i$$
$$\mathcal{D} := XS^\top + \mathcal{D}$$
$$\mathcal{D}^{(j)} := \mathcal{D}^{(j)} / ||\mathcal{D}^{(j)}||_2 \forall j$$

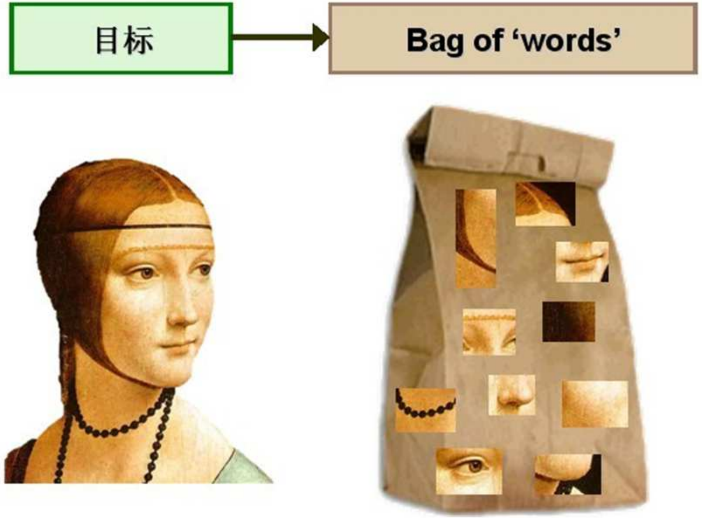


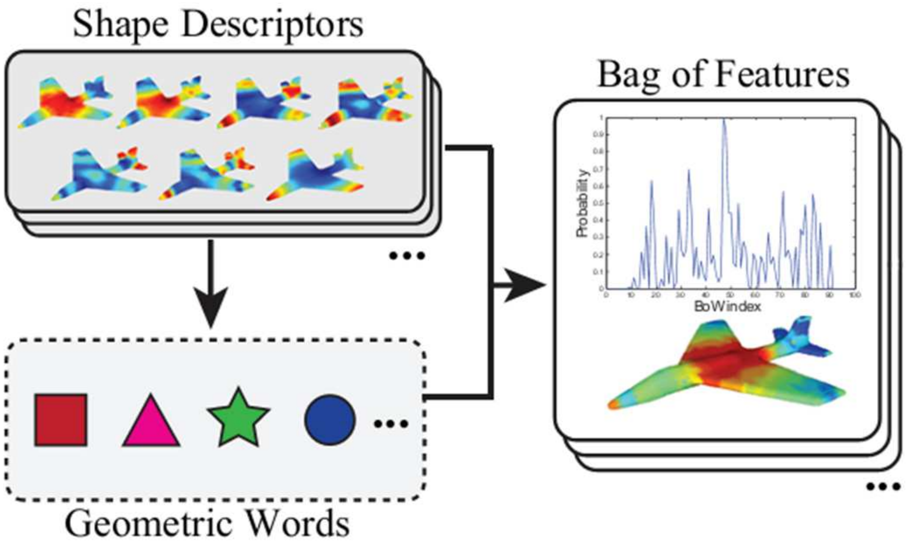


Adam Coates and Andrew Y. Ng, Learning Feature Representations with K-Means

Application: GIF

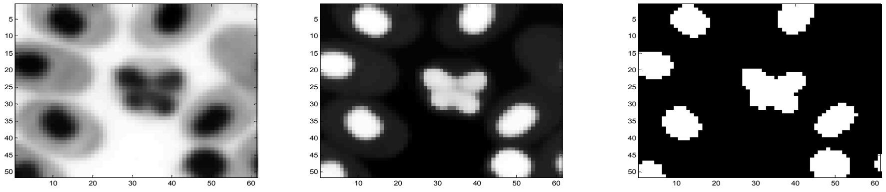
Application: Bag of Words





Application

- Bag of Words (BoW)
- Image segmentation
- Superpixel



Summary

- *K*-means algorithm is a simple yet popular method for clustering analysis
- Its performance is determined by initialisation and appropriate distance measure
- There are several variants of *K*-means to overcome its weaknesses
  - *K*-Medoids: resistance to noise and/or outliers
  - *K*-Modes: extension to categorical data clustering analysis
  - CLARA: extension to deal with large data sets
  - Mixture models (EM algorithm): handling uncertainty of clusters