lecture 7: clustering algorithms

STAT 545: Intro. to Computational Statistics

Vinayak Rao

Purdue University

September 12, 2016

Clustering

Given a large dataset, group data points into 'clusters'.

Data points in the same cluster are similar in some sense.

E.g. cluster students scores (to decide grade)

Applications:

Compression/feature-extraction/exploration/visualization

- simpler representation of complex data

Image segmentation, community detectn, co-expressed genes

Clustering (contd.)

We are given N data vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ in $\Re^d$.
Let c_i be the cluster assignment of observation x_i:

$$c_i \in \{1 \cdots K\}$$

Equivalently, we can use one-hot (or 1-of-K) encoding:

$$r_{ic} = \begin{cases} 1, & \text{if } c_i = c \\ 0, & \text{otherwise} \end{cases}$$

Observe: $r_{ic} \geq 0$ and $\sum_{c=1}^{K} r_{ic} = 1$ just like a probability vector.

However, r_{ic} is binary: we will relax this in later lectures.

Cluster parameters

Associate cluster i with parameter $\theta_i \in \Re^d$ (cluster prototype).

Write $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_K\}$, $\mathsf{C} = \{c_1, \ldots, c_N\}$ (or $\mathsf{R} = \{\mathsf{r}, \ldots, \mathsf{r}_N\}$).

Problem: Given data $(\mathsf{x}_1, \ldots, \mathsf{x}_N)$ find $\boldsymbol{\theta}$ and C.

Define a loss-function $L(\boldsymbol{\theta}, \mathsf{C})$, and minimize it.

Start by defining a distance (or similarity measure) $d(\mathsf{x}, \boldsymbol{\theta})$:

$$d(\mathsf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{d} (x_i - \theta_i)^2 \qquad \text{Squared Euclidean or } L_2 \text{ dist.}$$

$$d(\mathsf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{d} |x_i - \theta_i| \qquad L_1 \text{ distance}$$

Clustering loss function

We want all members of a cluster to be close to the prototype.

$\sum_{i \text{ s.t. } c_i = c} d(\mathbf{x}_i, \boldsymbol{\theta}_c) = \sum_{i=1}^{N} r_{ic} d(\mathbf{x}_i, \boldsymbol{\theta}_c)$ should be small for each c.

Overall loss function:

$$L(\boldsymbol{\theta}, \mathbf{R}) = \sum_{c=1}^{K} \sum_{i=1}^{N} r_{ic} d(\mathbf{x}_i, \boldsymbol{\theta}_c)$$

Optimize over both:

- cluster assignments (discrete)
- cluster parameters (continuous)

Minimizing $L(\boldsymbol{\theta}, \mathsf{C})$ is hard ($O(N^{DK+1})$).

Instead, use heuristic greedy (local-search) algorithms.
When $d(\cdot, \cdot)$ is Euclidean, the most popular is Lloyd's algorithm.

If we had the cluster parameters $\boldsymbol{\theta}^*$, can we solve for C?

$$\mathsf{C}_{opt} = \text{argmin } L(\boldsymbol{\theta}^*, \mathsf{C})$$

If we had the cluster assignments C^*, can we solve for $\boldsymbol{\theta}$?

$$\boldsymbol{\theta}_{opt} = \text{argmin } L(\boldsymbol{\theta}, \mathsf{C}^*)$$

K-means

Start with an initialialization of the parameters, call it $\boldsymbol{\theta}_0$.
Assign observations to nearest clusters, giving R_0.

Repeat for i in 1 to N:

- Recalculate cluster means, $\boldsymbol{\theta}_i$
- Recalculate cluster assignments, R_i

Coordinate-descent.

Resulting algorithm has complexity $O(INKD)$

[demo]

Questions

Does this algorithm converge to a global minimum?

Does it converge at all?

What is the convergence criteria?

Limitations

Local optima: Sensitive to initialization.
Solution: Run many times and pick the best clustering.

Empty clusters.
Solution: discard them, or use heuristics to assign observations to them

Choosing K.
Solution: search over a set of K's, penalizing larger values.

Requires circular clusters.
Solution: use some other method

Variations to k-means

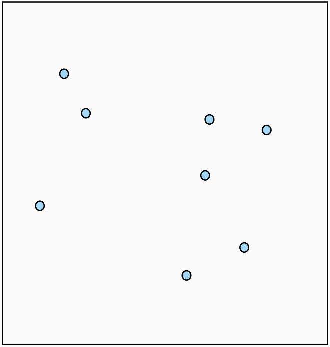
Modify distance functions.

L_1 distance: k-medians

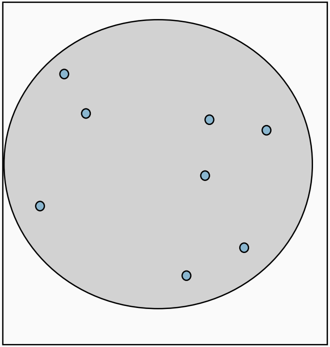
Modify the algorithm.

L_1 distance: k-medoids (exemplar-based)

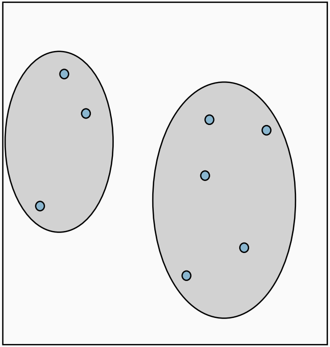
k-means is a partitioning algorithm that assigns each observation to a unique cluster.
Often there is no clear best clustering.



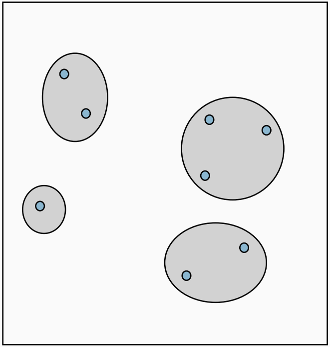
k-means is a partitioning algorithm that assigns each observation to a unique cluster.
Often there is no clear best clustering.



k-means is a partitioning algorithm that assigns each observation to a unique cluster.
Often there is no clear best clustering.

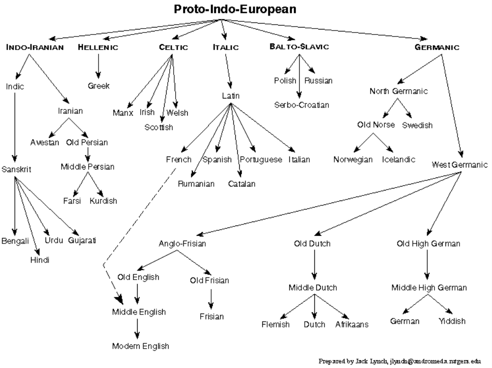


k-means is a partitioning algorithm that assigns each observation to a unique cluster.
Often there is no clear best clustering.



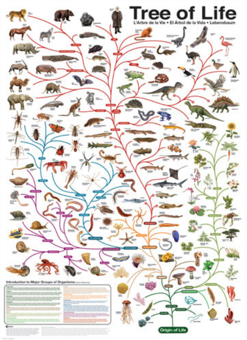
k-means is a partitioning algorithm that assigns each observation to a unique cluster.

Often is is natural to view data as having a hierarchical structure



Hierarchical clustering

k-means is a partitioning algorithm that assigns each observation to a unique cluster.
Often is is natural to view data as having a hierarchical structure



Two approaches:

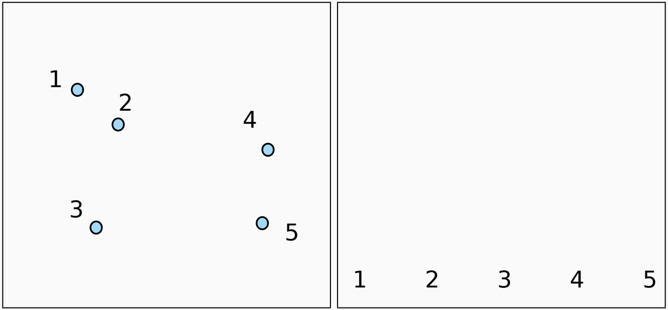
Top-down (divisive) clustering:

- Initialize all observations into a single cluster, and divide clusters sequentially.

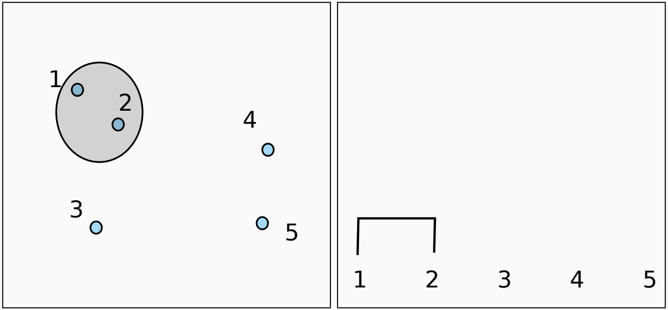
Bottom-up (agglomerative) clustering:

- Initialize each observation in its own cluster, and merge clusters sequentially.
- More flexible, and more common.

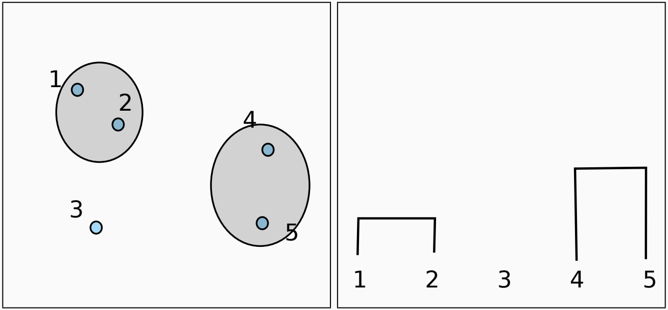
Agglomerative clustering

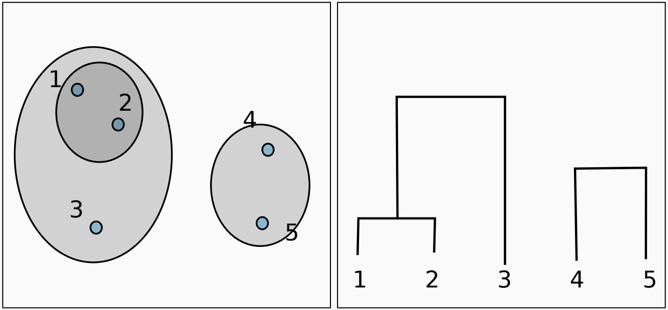


Agglomerative clustering

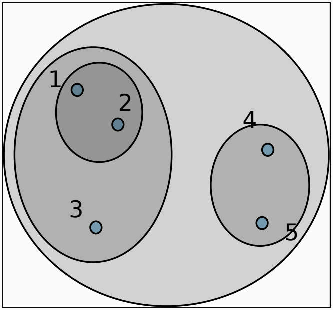


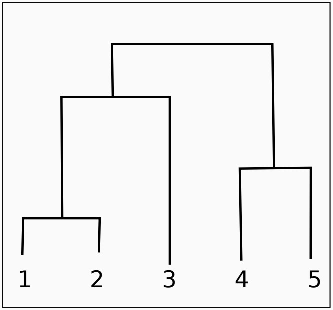
Agglomerative clustering

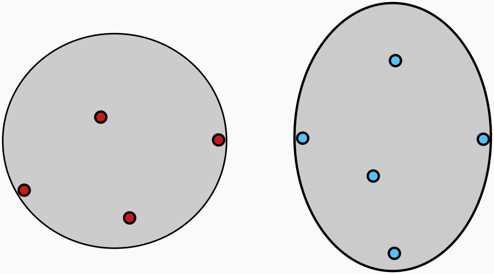



1
2
3
4
5
1
2
3
4
5

Agglomerative clustering



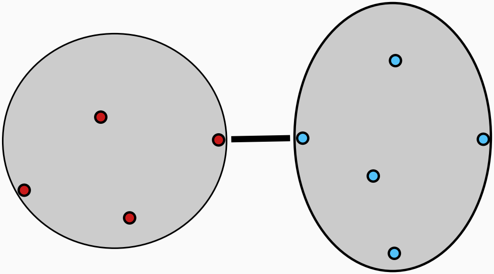




Pick a distance function (e.g. Euclidean).

Pick a linkage criterion defining distance between two clusters:

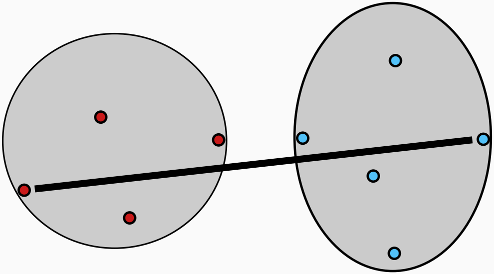
- Single linkage: $d(A, B) = \min_{x \in A, y \in B} d(x, y)$.
- Complete linkage: $d(A, B) = \max_{x \in A, y \in B} d(x, y)$.
- Centroid linkage: $d(A, B) = d(C_A, C_B)$ (C_A: centroid of A).
- Average linkage: $d(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y)$.



Pick a distance function (e.g. Euclidean).

Pick a linkage criterion defining distance between two clusters:

- Single linkage: $d(A, B) = \min_{x \in A, y \in B} d(x, y)$.
- Complete linkage: $d(A, B) = \max_{x \in A, y \in B} d(x, y)$.
- Centroid linkage: $d(A, B) = d(C_A, C_B)$ (C_A: centroid of A).
- Average linkage: $d(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y)$.

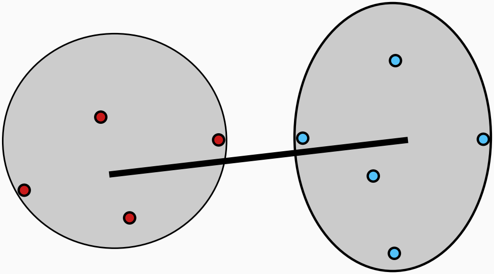


Pick a distance function (e.g. Euclidean).

Pick a linkage criterion defining distance between two clusters:

- Single linkage: $d(A, B) = \min_{x \in A, y \in B} d(x, y)$.
- Complete linkage: $d(A, B) = \max_{x \in A, y \in B} d(x, y)$.
- Centroid linkage: $d(A, B) = d(C_A, C_B)$ (C_A: centroid of A).
- Average linkage: $d(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y)$.

Agglomerative clustering

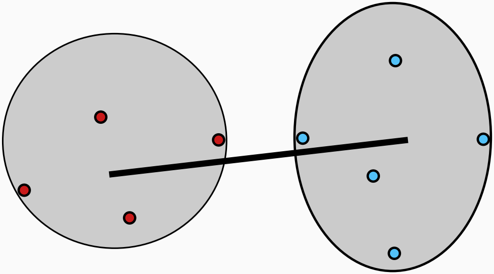


Pick a distance function (e.g. Euclidean).

Pick a linkage criterion defining distance between two clusters:

- Single linkage: $d(A, B) = \min_{x \in A, y \in B} d(x, y)$.
- Complete linkage: $d(A, B) = \max_{x \in A, y \in B} d(x, y)$.
- Centroid linkage: $d(A, B) = d(C_A, C_B)$ (C_A: centroid of A).
- Average linkage: $d(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y)$.

Agglomerative clustering

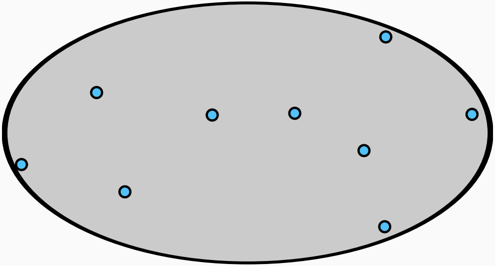


Pick a distance function (e.g. Euclidean).

Pick a linkage criterion defining distance between two clusters:

- Single linkage: $d(A, B) = \min_{x \in A, y \in B} d(x, y)$.
- Complete linkage: $d(A, B) = \max_{x \in A, y \in B} d(x, y)$.
- Centroid linkage: $d(A, B) = d(C_A, C_B)$ (C_A: centroid of A).
- Average linkage: $d(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y)$.

Agglomerative clustering



Pick a distance function (e.g. Euclidean).

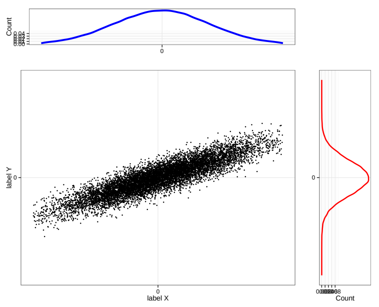
Pick a linkage criterion defining distance between two clusters:

- Single linkage: $d(A, B) = \min_{x \in A, y \in B} d(x, y)$.
- Complete linkage: $d(A, B) = \max_{x \in A, y \in B} d(x, y)$.
- Centroid linkage: $d(A, B) = d(C_A, C_B)$ (C_A: centroid of A).
- Average linkage: $d(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y)$.

Some properties of the Gaussian

Marginalization:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right), \quad X \sim ?$$

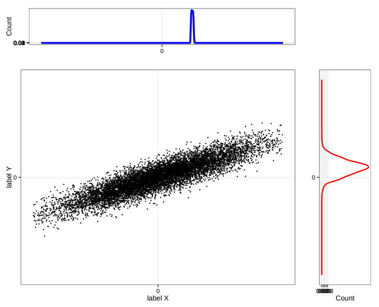


$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 10 & 4 \\ 4 & 2 \end{bmatrix}$$

$$X \sim \mathcal{N}\left(\mu_X, \Sigma_{XX}\right)$$

Conditioning:

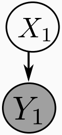
$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right), \quad Y|(X = a) \sim ?$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 10 & 4 \\ 4 & 2 \end{bmatrix}$$

$$Y|X \sim \mathcal{N}\left(\mu_Y + \Sigma_{XY}\Sigma_{XX}^{-1}(a - \mu_X)\ ,\ \Sigma_{YY} - \Sigma_{XY}\Sigma_{XX}^{-1}\Sigma_{YX}\right)$$

The Gaussian distribution, conjugacy and Bayes' rule



$$X_1 \sim \mathcal{N}(\mu, \Sigma)$$

$$\epsilon \sim \mathcal{N}(0, \Sigma_E)$$

$$Y_1 = AX_1 + \epsilon$$

We have a Gaussian 'prior' on X_1.
We observe a noisy measurement $Y_1|X_1 \sim \mathcal{N}(AX_1, \Sigma_E)$.

$$\begin{bmatrix} X \\ \epsilon \end{bmatrix} \to \begin{bmatrix} X \\ Y \end{bmatrix}$$

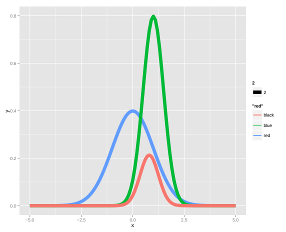
X and Y jointly Gaussian: what is its mean and covariance?

Y is marginally Gaussian: what is its mean and covariance?

$X|Y$ is Gaussian: what is its mean and covariance?

Product of Gaussian densities:

Product of Gaussian densities is Gaussian (upto normalization)



Intuition: sum of two quadratic functions is a quadratic

Aside: need only specify prob. distrib. to a constant

- $p(x)$ and $C \cdot p(x)$ represents the same, if C is independent of x
- Probabilities must integrate to 1