Date: Tuesday, February 19, 2019.

Clustering

↳ Task of unsupervised learning (just data, no labels)

Aim: Take data points (vectors) $X_1, \ldots, X_n$ and identify groups of closely related points. These groups, or clusters, generally contain points that are "close" to one another but "far" from points outside their cluster.

Note:

- The choice of distance greatly affects the outcome.
- Estimating the number of clusters can be challenging but generally heuristics like elbow plots or the spectral gap can be used.

"Off-the shelf" clustering methods

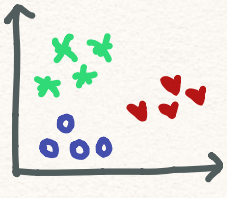
1) K-means:

- Identify K clusters whose points have minimum within-cluster sum of squared distance

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{j=1}^{K} \sum_{i=1}^{n} (X_i - C_j)^2 \, 1\{\text{data point } i \text{ is in cluster } j\} \right\}$$

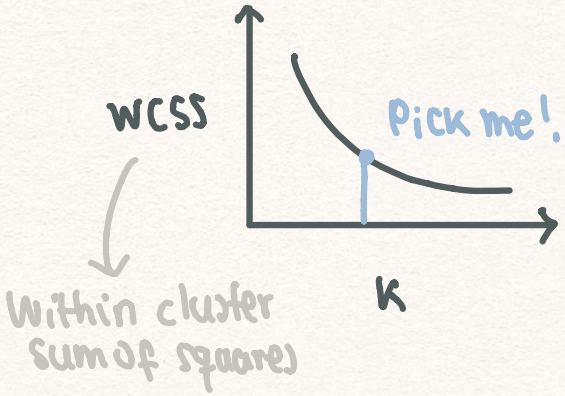
Here C_j = the mean of cluster j.

- This leads to spherical looking clusters.



- K-means can easily be extended to medians rather than means, this is called K-medoids.

- Choice of k often relies on an elbow plot



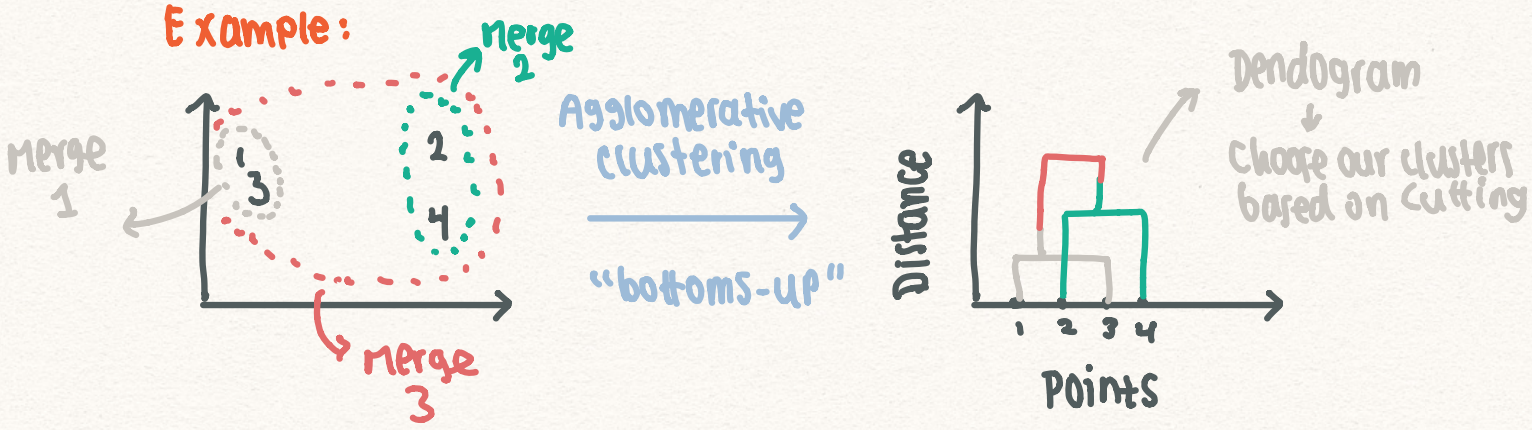
2) Hierarchical clustering:

- Tree-based clustering algorithm that seeks clusters with minimum within cluster distance.
- Distances of ① point to point and ② clustering to point are chosen by the user.

Dissimilarity metric

Linkage

Example:



3) Spectral clustering:

- Motivated by non-spherical clusters that share some basic shape/connectivity
- It is a graph based method
- Pseudo-code (with no further description)
  1) Takes the similarity matrix of points and calculates the graph Laplacian
  2) Performs eigen-decomposition of the Laplacian

3) Stacks K smallest eigenvectors side by side in a matrix X.

4) Clusters rows of X using K-means.

For more info, see Von Luxbourg's "A Tutorial on Spectral clustering".

* The most common clustering algorithms are not based on any model of the data.
* There are model-based clustering methods which assume that the data come from a mixture of distributions.
* When looking at a histogram of data, if we see bimodal (or polymodal) curves, we may believe that a **mixture model**, i.e. a model where data comes from some mixture of distributions, best fits the data.

Common example:

Gaussian mixture model:

$$X_i \sim pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2)$$

Here, p = probability of coming from the $N(\mu_1, \sigma_1^2)$.

Aim: Estimate $p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2$

Clustering objective: Estimate the probability of each data point x_i coming from cluster 1 or 2.

Let C_i = the cluster for data point i.

Aim: Calculate $P(C_i | \underset{\sim}{X})$

Prior: $C_i \sim$ Discrete uniform(1,2)

↳ This gives equal probabilities to both clusters but we know the $P(C_i = 1) = p$

So, we incorporate this and draw C_i from a Categorical distribution w/ probabilities $(p, 1-p)$ as our prior.

Pycm2

Precision: $\frac{1}{\sigma^2} = \tau$

We have $C_1, C_2, \ldots C_n$

Priors on other parameters: $\sigma_1^2 \sim \text{Uniform}(0, 100)$

$\sigma_2^2 \sim \text{Uniform}(0, 100)$

Reasonable from looking at the histogram

$\mu_1 \sim N(120, 100)$

$\mu_2 \sim N(190, 100)$

We have priors for $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, c_1, \ldots, c_n$ and the data generating process for $x_1, \ldots, x_n$ (Gaussian mixture model). So now we can use MCMC to calculate posterior for each parameter.

Observed data

About the exam: → Chapter 3 + Slides

- Is pseudo-code correct?
- What algorithm is it?
- Calculate values based on the pseudo-code.
- Diagnostic plots for MCMC
- Understand how to model a Markov Chain
- Bayesian Clustering
- Page rank + Markov Chain