University of Colorado **Boulder**

CSCI 5622: Machine Learning

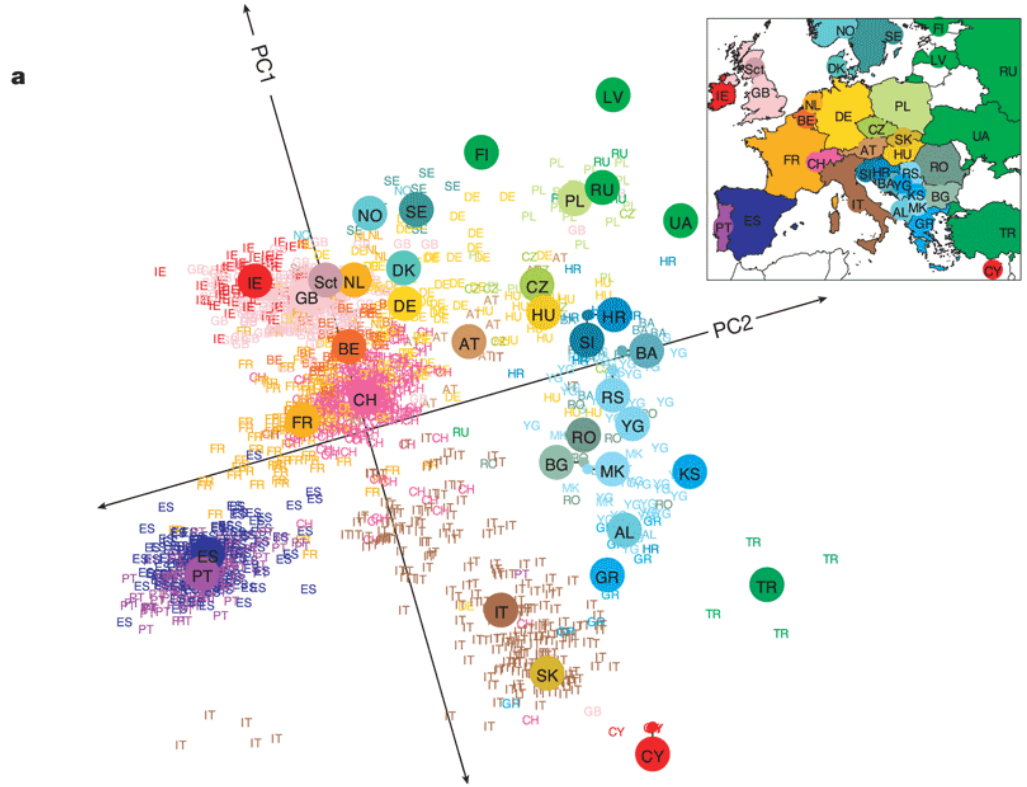
Lecture 12

Overview

- Clustering overview
- Partitional clustering
  - K-means clustering
  - Gaussian Mixture Models (GMM)

Clustering

(1) Understanding: Finding patterns/structure/sub-populations in data



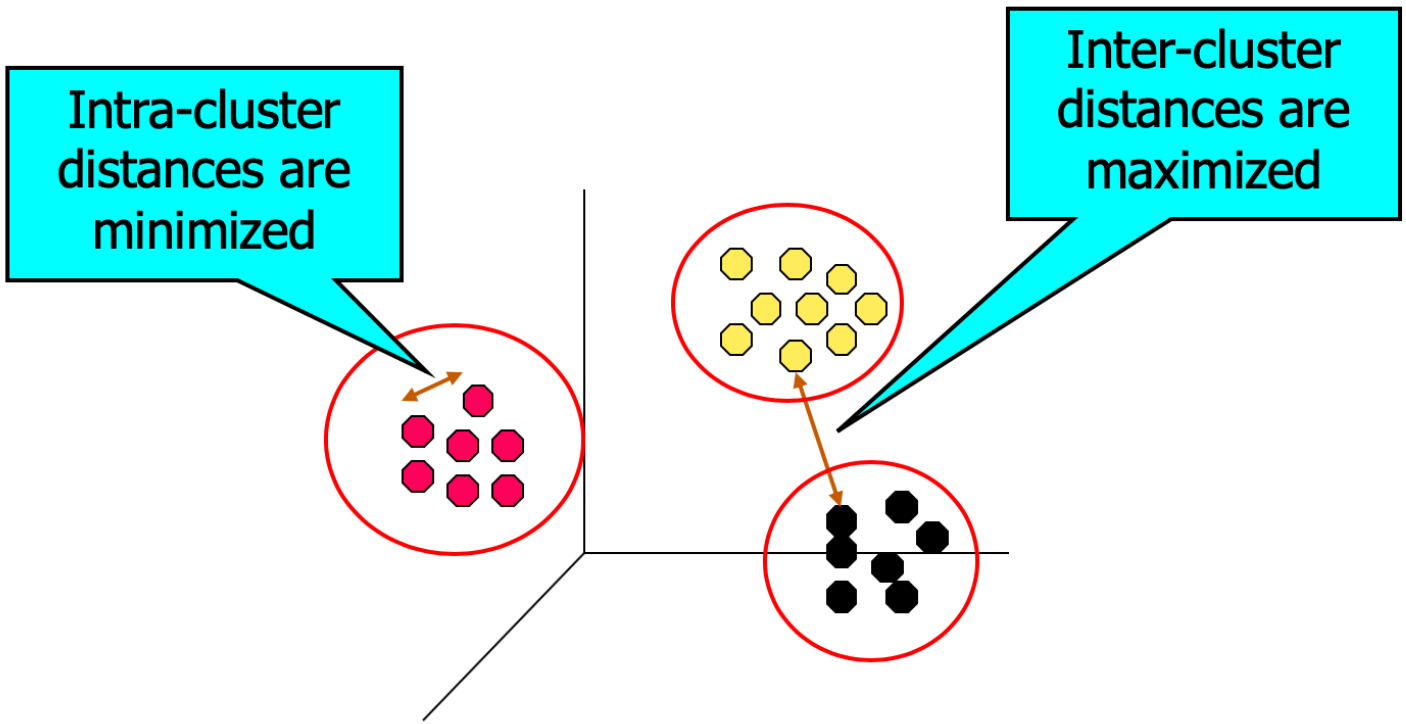
(2) Summarization: Reducing the size of large datasets

Clustering

- find patterns/structure/sub-populations in data ("knowledge discovery")
- training data does not include desired outputs
- less well-defined problem with no obvious error metrics
- topic modeling, market segmentation, clustering of hand-written digits, news clustering (e.g. Google news)

Clustering

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



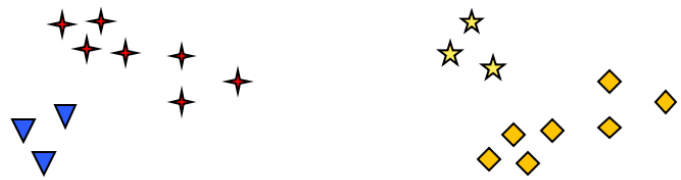
Clustering

Notion of clustering can be ambiguous

How many clusters?

Six Clusters

Two Clusters



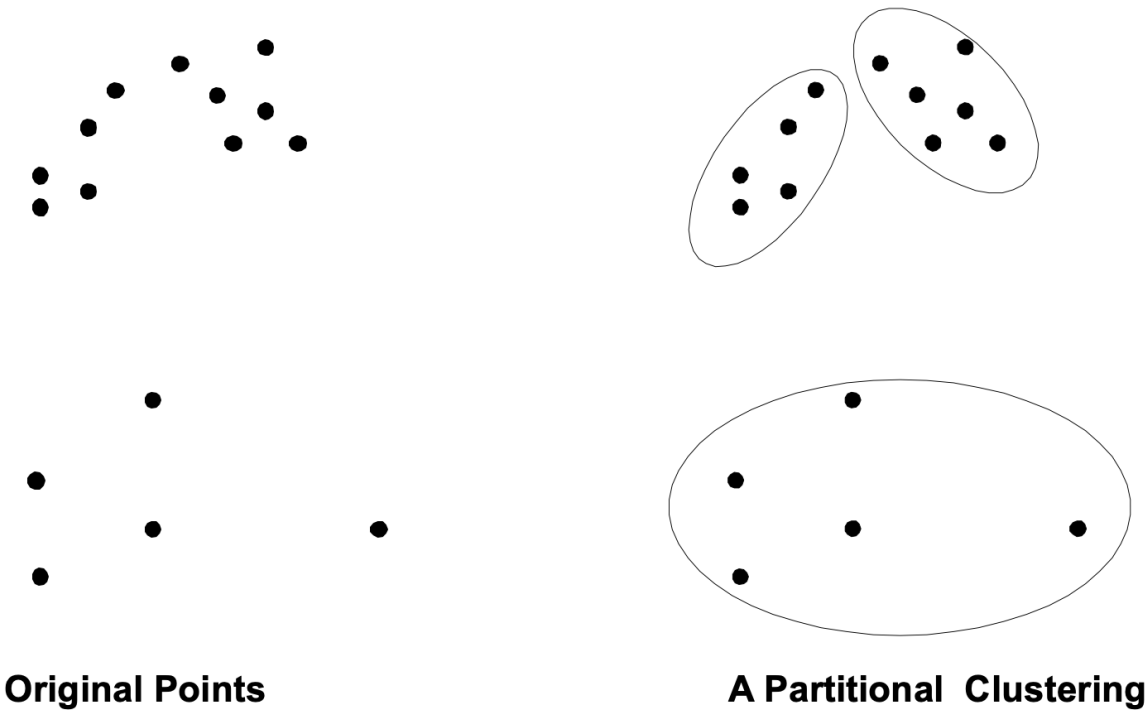
Four Clusters

Types of clustering

- Partitional clustering
  - non-hierarchical clusters
- Hierarchical clustering
  - a set of nested clusters organized as a hierarchical tree

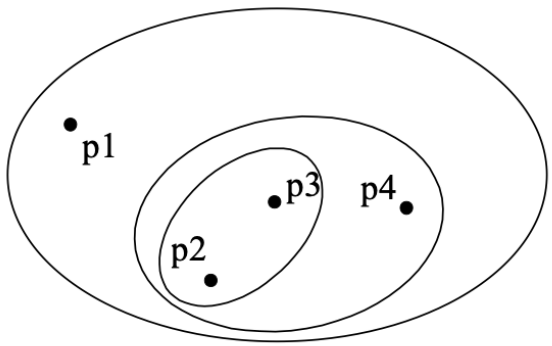
Types of clustering

Partitional clustering

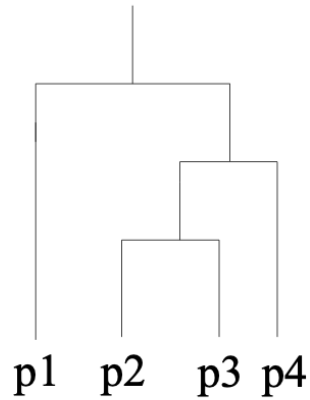


Types of clustering

Hierarchical clustering



Traditional Hierarchical Clustering



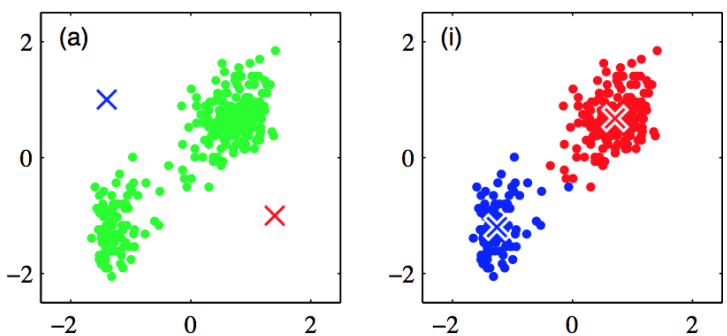
Traditional Dendrogram

Overview

- Clustering overview
- Partitional clustering
  - K-means clustering
  - Gaussian Mixture Models (GMM)

K-means Clustering: Representation

- Input: Data $\mathcal{D} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$
- Output: Clusters $\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}$
- Decision: Cluster membership, the cluster id assigned to sample $\mathbf{x_n}$, i.e. $A(\mathbf{x_n}) \in \{1, \ldots, K\}$
- Evaluation metric: Distortion measure
  $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x_n} - \boldsymbol{\mu_k}\|_2^2$, where $r_{nk} = 1$ if sample n is assigned to cluster k ($A(\mathbf{x_n}) = k$), 0 otherwise
- Intuition: Data points assigned to cluster k should be close to centroid $\boldsymbol{\mu_k}$



K-means Clustering

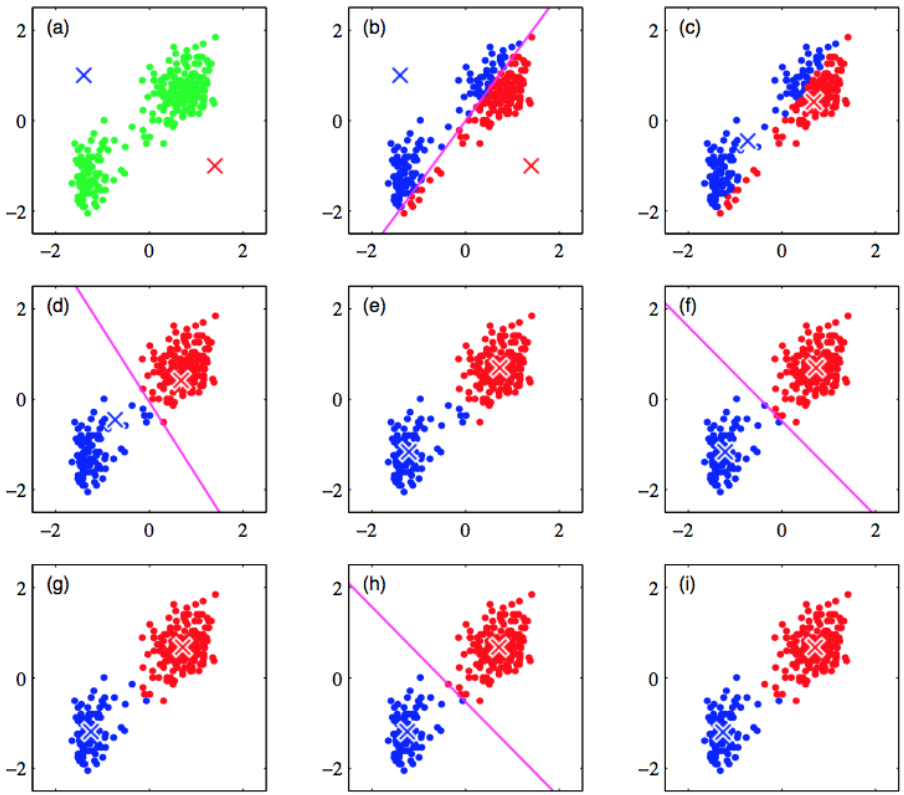
Evaluation metric: $\min_{r_{nk}} J = \min_{r_{nk}} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x_n} - \boldsymbol{\mu_k}\|_2^2$

Optimization:

- Step 0: Initialize $\boldsymbol{\mu_k}$ to some values
- Step 1: Assume the current value of $\boldsymbol{\mu_k}$ fixed, minimize J over r_{nk}, which leads to the following cluster assignment rule
  $r_{nk} = \begin{cases} 1, & \text{if } k = arg\min_j \|\mathbf{x_n} - \boldsymbol{\mu_j}\|_2^2 \\ 0, & \text{otherwise} \end{cases}$
- Step 2: Assume the current value of r_{nk} fixed, minimize J over $\boldsymbol{\mu_k}$, which leads to the following rule to update the prototypes of the clusters $\boldsymbol{\mu_k} = \frac{\sum_n r_{nk} \mathbf{x_n}}{\sum_n r_{nk}}$
- Step 3: Determine whether to stop or return to Step 1

K-means Clustering

Example



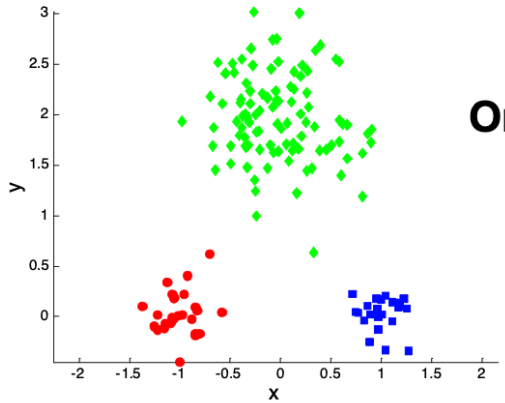
K-means Clustering

Remarks

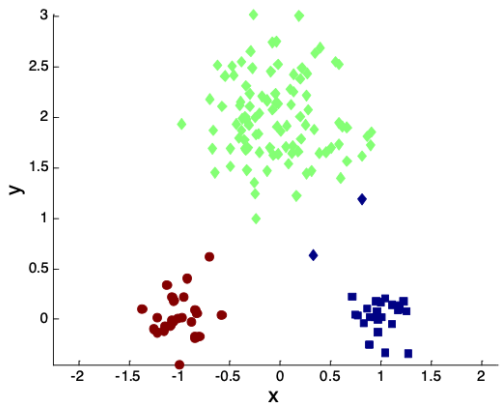
- The centroid $\boldsymbol{\mu_k}$ is the means of data points assigned to the cluster k, hence the name K-means clustering.
- The procedure terminates after a finite number of steps, as the procedure reduces J in both Step 1 and Step 2
- There is no guarantee the procedure terminates at the global optimum of J. In most cases, the algorithm stops at a local optimum, which depends on the initial values in Step 0 $\rightarrow$ random restarts to improve chances of getting closer to global optima

K-means Clustering

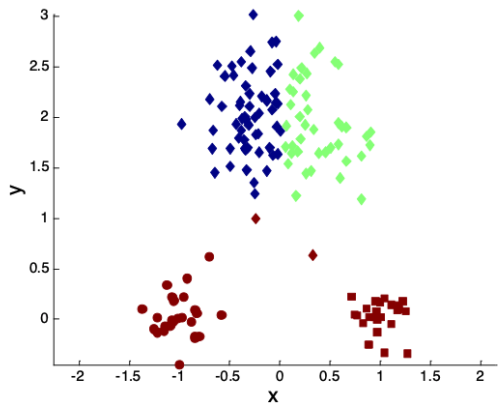
Initialization of K-Means is important



Original Points



Optimal Clustering



Sub-optimal Clustering

K-means Clustering

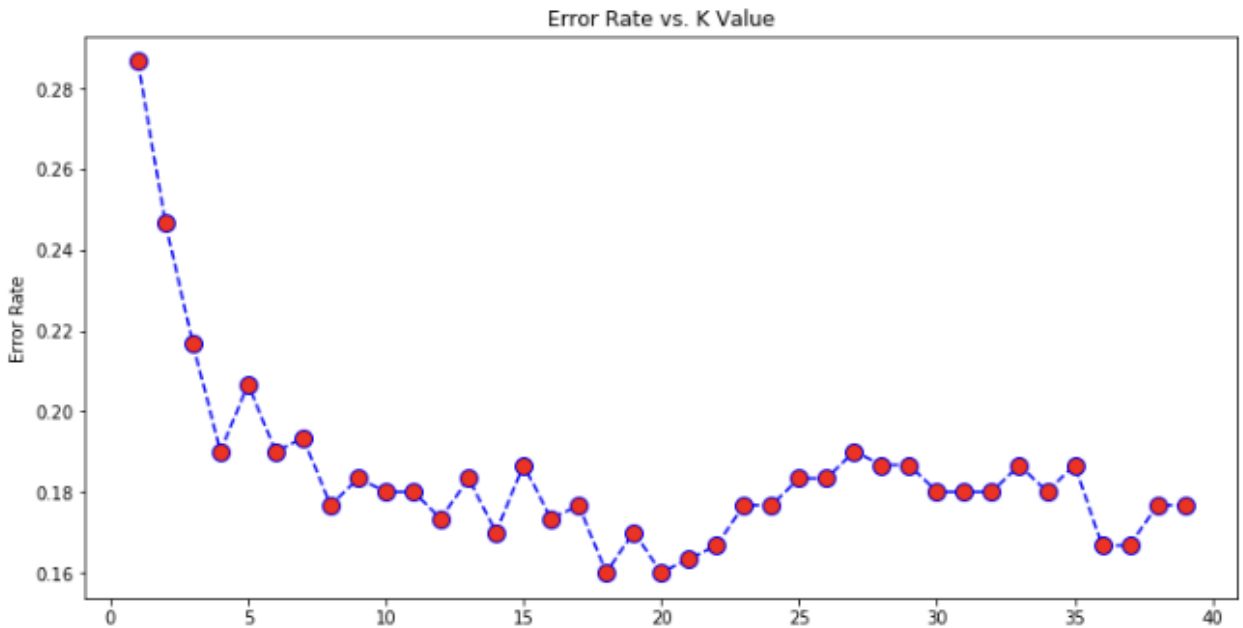
Solutions to Initial Centroids Problem

- Multiple random initializations
- Start with hierarchical clustering to determine initial centroids
- Select more than K initial centroids and then select among these initial centroids

K-means Clustering

How to know when to stop - Elbow Method

- Plot the error (i.e., distance of each sample to the corresponding centroid) against the number of clusters
- Stop when the decrease in error becomes almost flat



K-means Clustering

Application: vector quantization

- We can replace our data points with the centroids $\boldsymbol{\mu_k}$ from the clusters they are assigned to $\rightarrow$ vector quantization
- We have compressed the data points into
  - a codebook of all the centroids $\{\boldsymbol{\mu_1}, \ldots, \boldsymbol{\mu_K}\}$
  - a list of indices to the codebook for the data points (created based on r_{nk})
- This compression is obviously lossy as certain information will be lost if we use a very small K

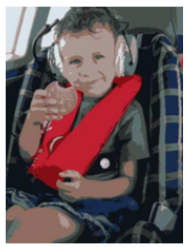
K-means Clustering

Question: vector quantization with K-means

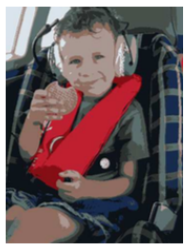
Assume that the images below are created by vectoring the original image with K-means using different values of K. What is the correct combination?

Original Image

A) $K = 25$ $K = 10$ $K = 3$



B) $K = 3$ $K = 10$ $K = 25$



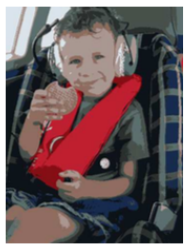
K-means Clustering

Question: vector quantization with K-means

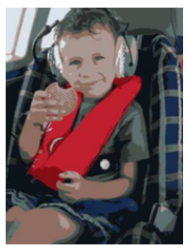
Assume that the images below are created by vectoring the original image with K-means using different values of K. What is the correct combination?

Original Image

A) $K = 25$ $K = 10$ $K = 3$



B) $K = 3$ $K = 10$ $K = 25$



Correct answer is A

K-means Clustering

Limitations of K-Means

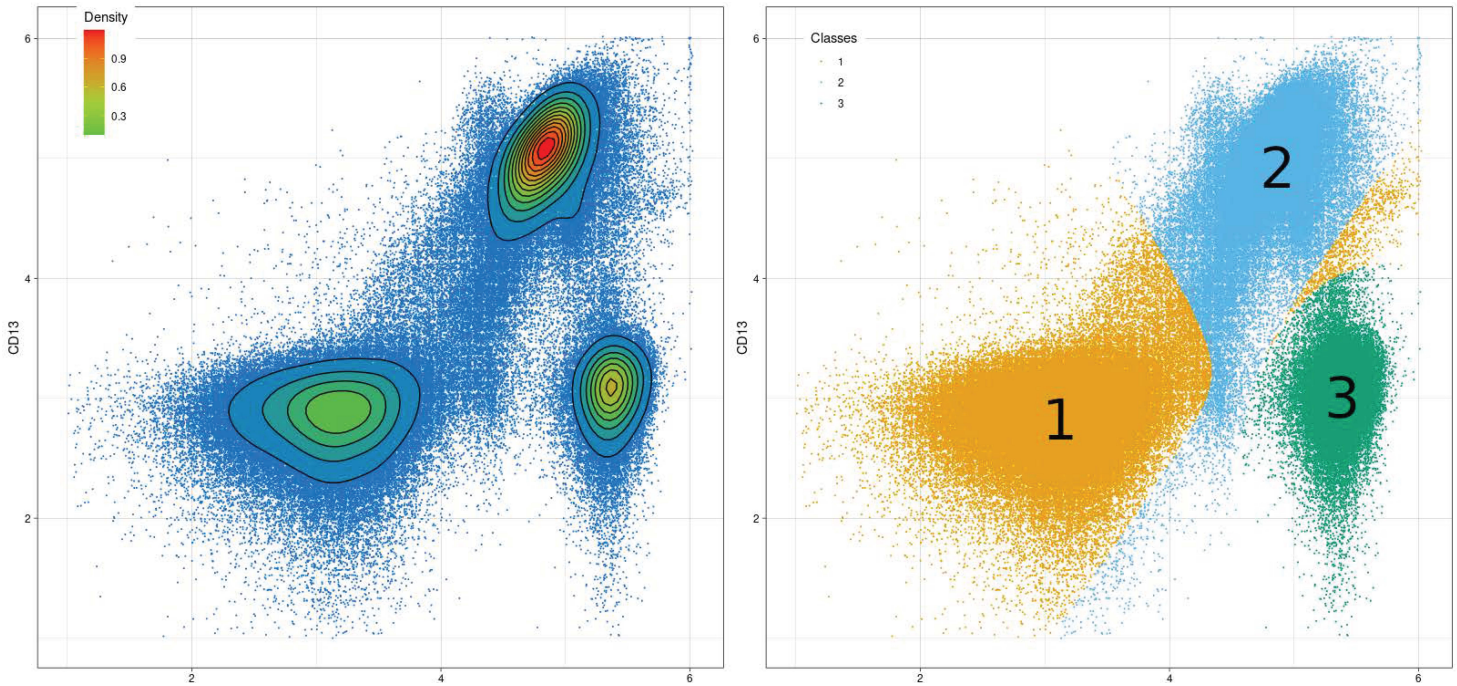
- Problems when clusters are of differing size, density, or non-spherical shapes (for Euclidean distances)
- Sensitive to outliers
- Number of clusters is difficult to determine

Overview

- Clustering overview
- Partitional clustering
  - K-means clustering
  - Gaussian Mixture Models (GMM)

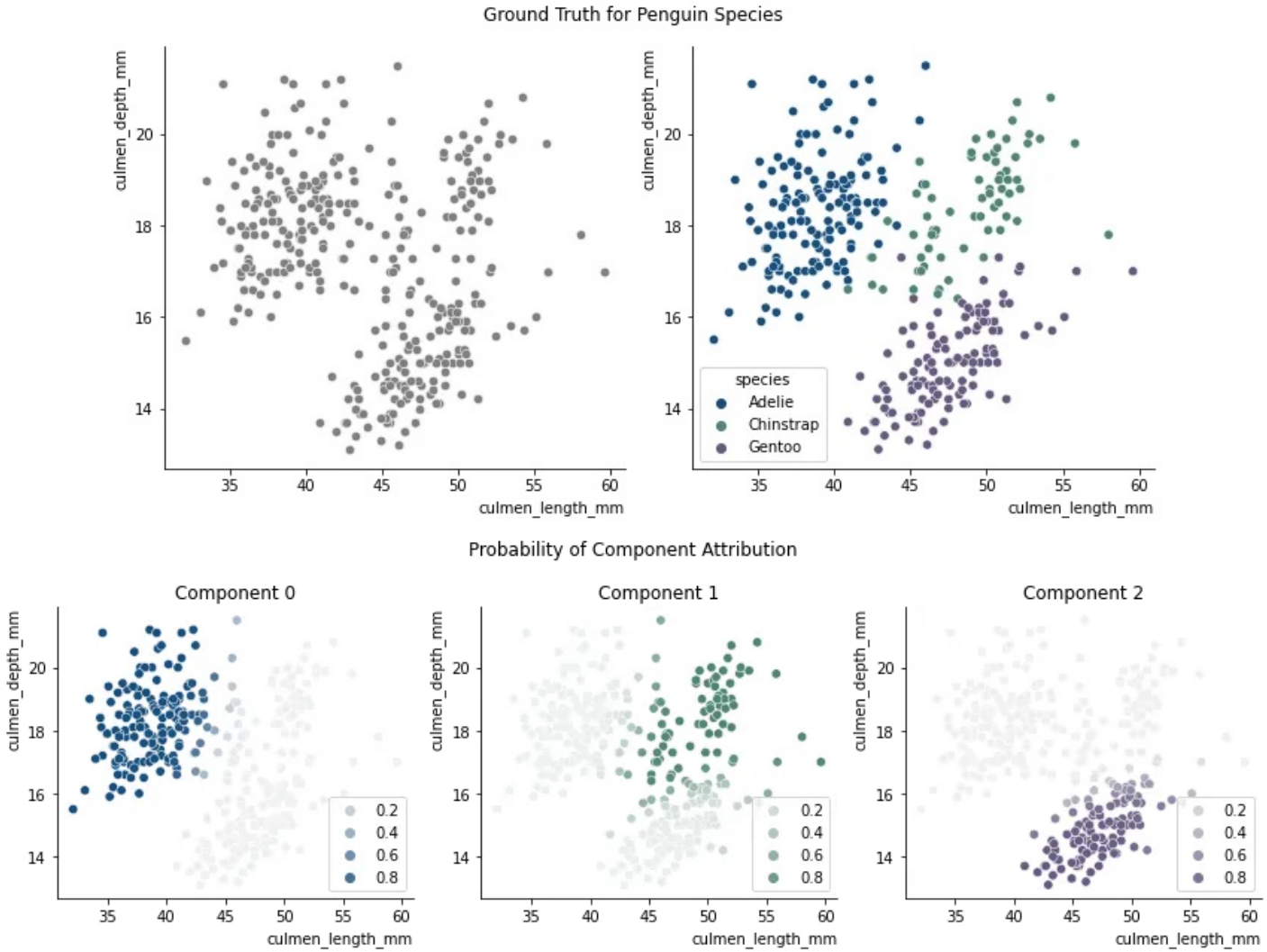
Gaussian Mixture Models: GMMs used for classification

Probabilistic interpretation of clustering



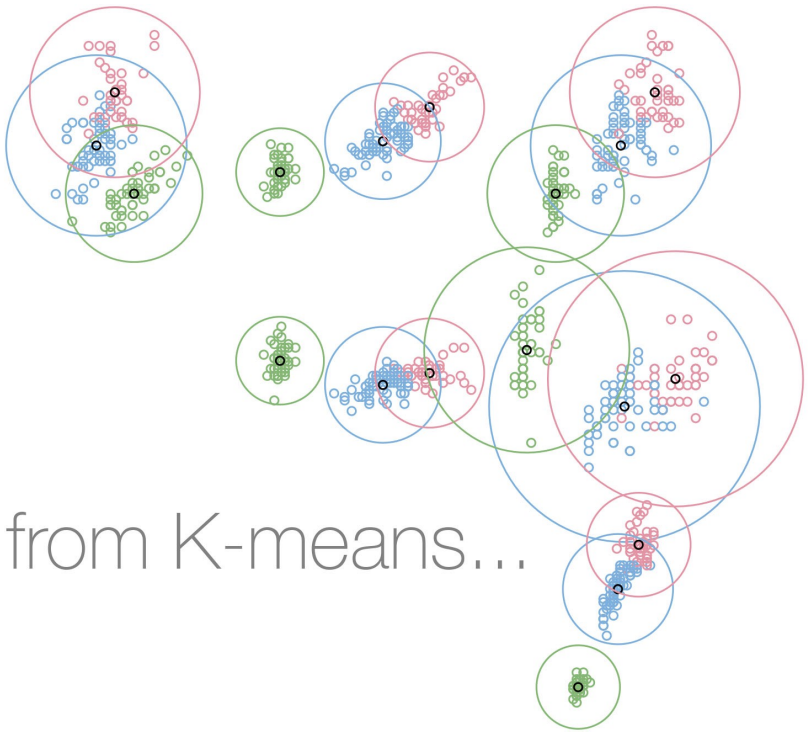
Gaussian Mixture Models: GMMs used for classification

Probabilistic interpretation of clustering



Gaussian Mixture Models

Comparison between K-Means and GMMs



from K-means...

... to GMM

Gaussian Mixture Models: Anomaly Detection

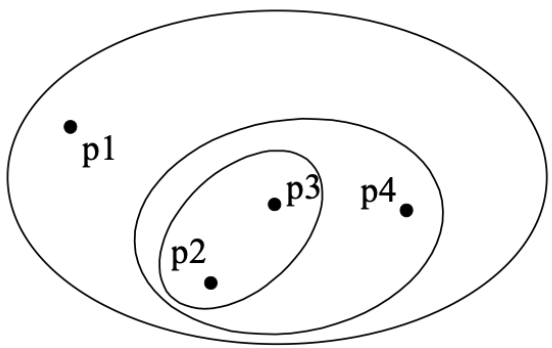
- GMMs can model 'normal' behavior in the data and identify instances that deviate from this behavior
- Detecting fraudulent credit card transactions (fraudulent transactions $<$ legitimate ones)
- Modeling the distribution of legitimate transactions and identifying samples that deviate from this distribution

Overview

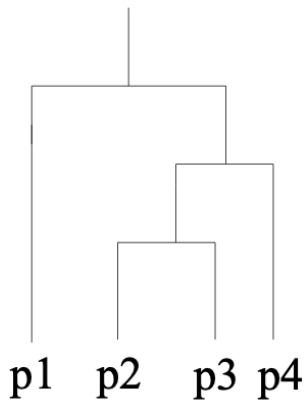
- Clustering overview
- Partitional clustering
  - K-means clustering
  - Gaussian Mixture Models (GMM)
- Hierarchical clustering

Hierarchical clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits



Traditional Hierarchical Clustering



Traditional Dendrogram

Hierarchical clustering

Advantages of hierarchical clustering

- Do not have to pre-determine number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level
- Resulting clusters may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction)

Hierarchical clustering

Types of hierarchical clustering

- Agglomerative
  - Start with each sample as individual cluster
  - Merge the closest pair of clusters each time until only one cluster left
- Divisive
  - Start with one, all-inclusive cluster
  - Split a cluster each time until each cluster contains a point

Overview

- Clustering tries to find patterns/hidden structures in data
- Partitional clustering
  - K-means: hard assignment of samples to one centroid
  - GMMs: soft assignment of samples to each Gaussian
- **Readings:** Alpaydin 7