Scientific Computing for Biologists

Lecture 10: K-Means Clustering and Mixture Models

Instructor: Paul M. Magwene

08 November 2011

Outline of Lecture

- K-means clustering
- Mixture model based clustering

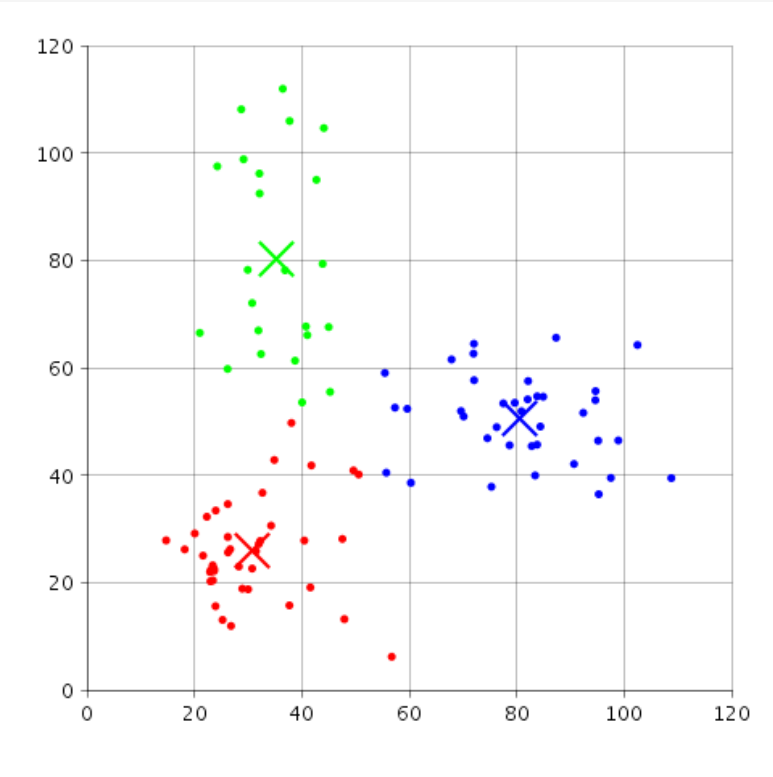
K-mean Clustering

General idea

Assign the n data points (or p variables) to one of K clusters to as to optimize some criterion of interest.

- The most common criterion to minimize is the sum-of-squares from the group centroids.

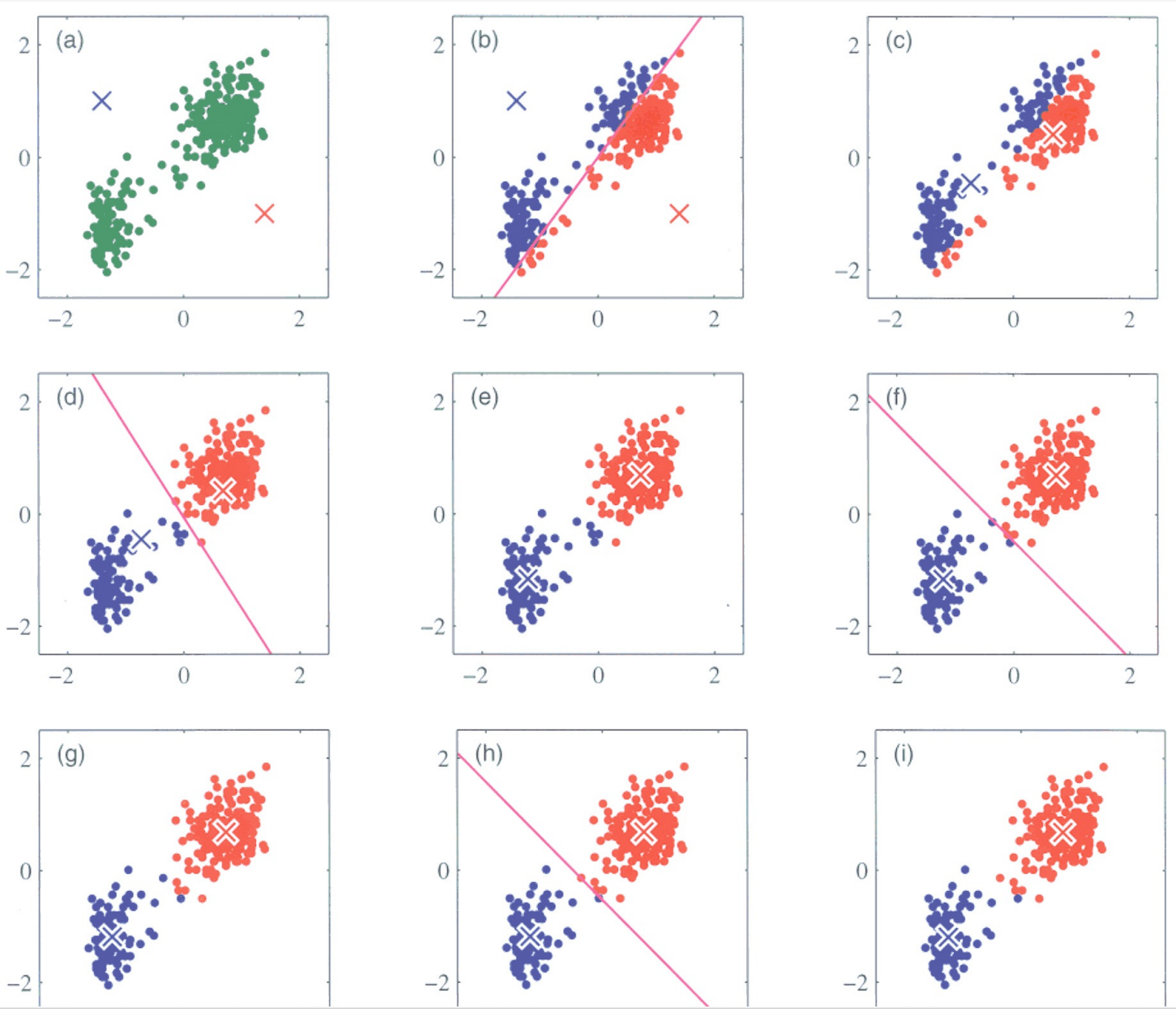
$$V = \sum_{i=1}^{k} \sum_{j \in g_i} |x_j - \mu_i|^2$$



Simple algorithm for K-means clustering

1. Decide on k, the number of groups
2. Randomly pick k of the objects to act as the initial centers
3. Assign each object to the group whose center it is closest to
4. Recalculate the k centers as the centroids of the objects assigned to them
5. Repeat from step 3 until centroids no longer move (convergence)

Illustration of K-means algorithm



Things to note re:K-means clustering

- The algorithm described above does not necessarily find the global optimum
- The algorithm is sensitive to choice of initial cluster center; k-means is often run multiple-time with different initial centers to insure inferred clusters are robust.

Clustering with Mixture Models

Goal

Method for assigning observations to clusters and estimating parametric distributions that describe the clusters.

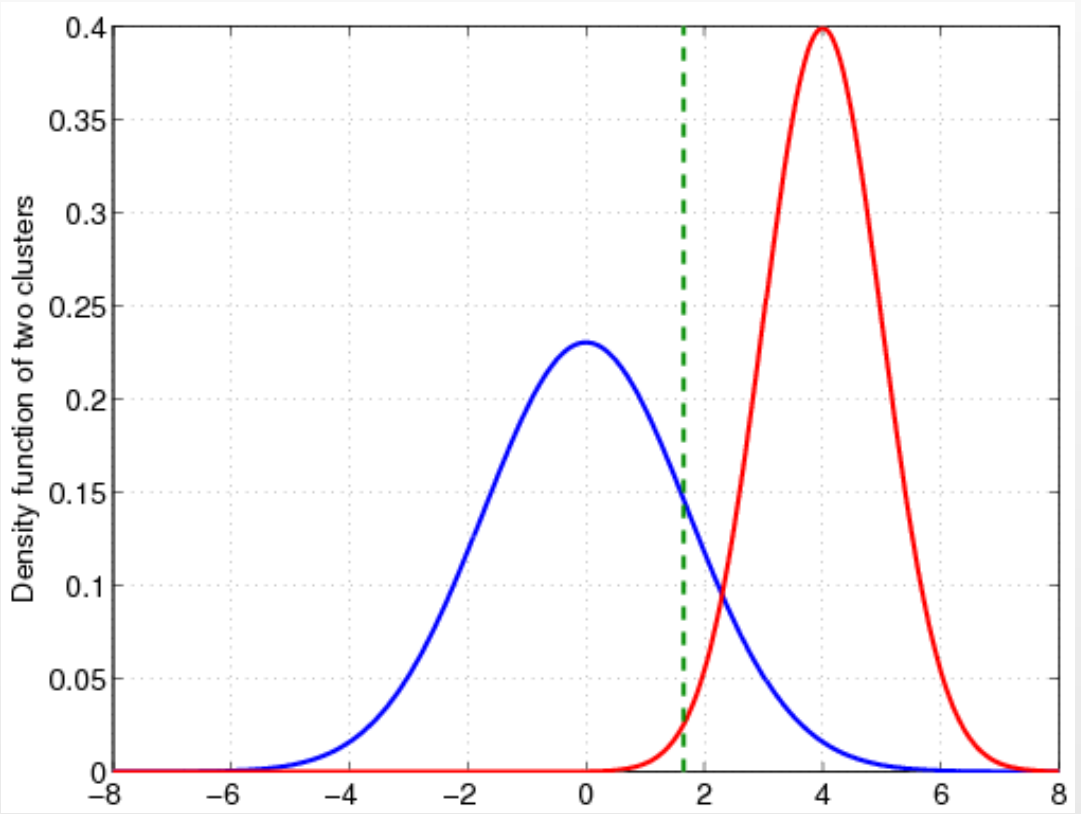
Assume that the data set represents observations drawn from a mixture of g sub-distributions (user specifies g), and that the probability density function of the mixture is given by:

$$p(\mathbf{x}) = \sum_{s=1}^{g} \pi_s p(\mathbf{x}; \theta_s)$$

Where the $p(\mathbf{x}; \theta_s)$ represents the s-th 'component density' (sub-distributions) and the θ_s are the component parameters. The π_s represent the weighting factor of the s-th component in the mixture.

Advantages

- Well-studied statistical inference techniques available.
- Flexibility in choosing the component distributions.
- Obtain a density estimation for each cluster.
- A "soft" classification is available.



Gaussian Mixture Models

A common starting point in mixture modeling is to assume that the components are Gaussian.

Therefore the components are multivariate Gaussian distributions:

$$N(\mathbf{x};\theta) \equiv (2\pi)^{-D/2}|\Sigma|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}-(\mathbf{x}-\mu)\right]$$

each with a different mean vector, μ ($\mu \in \mathbb{R}^p$), and covariance matrix, Σ ($p \times p$).

How do we 'solve' the mixture model problem?

The mixture model problem involves optimization over multiple parameters.

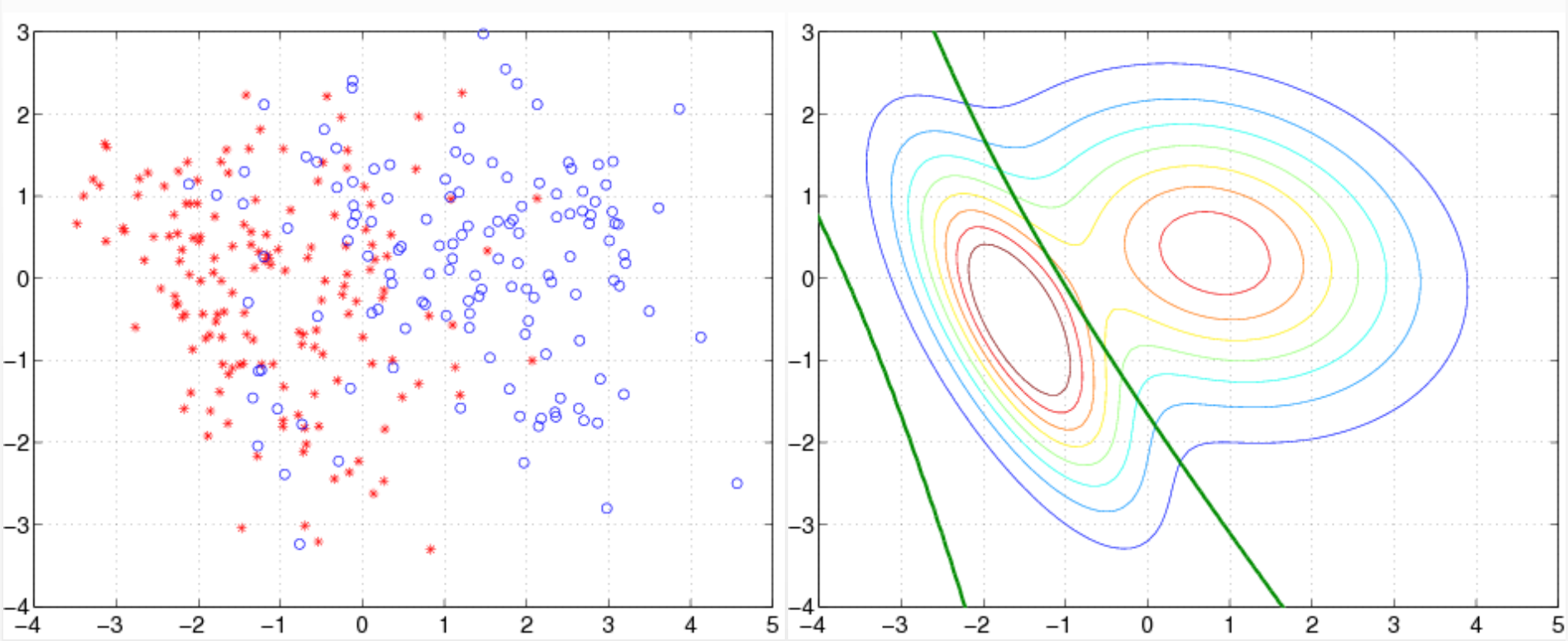
The standard approach to estimating the parameters is called the "Expectation-Maximization" (EM) algorithm.

- Described by Dempster, Laird, and Rubin (1977)
- Provides a way to iterative compute a maximum likelihood estimation when the observed data are incomplete or there are 'latent' parameters.

Overview of the EM Algorithm

1. Guess a set of starting parameters
2. Use these starting parameters to 'estimate' the complete data
3. Use the estimates of the complete data to update the parameters
4. Repeat steps 2 and 3 until convergence

Mixture Model Clustering, Example



Heart disease example: 297 samples (137 with heart disease). 13 quantitative varibles (e.g. cholesterol, max heart rate, etc). Data centered and normalized. Data projected onto first two PCs. Two-component Gaussian mixture fit.