Mixture Models

Chapter 8

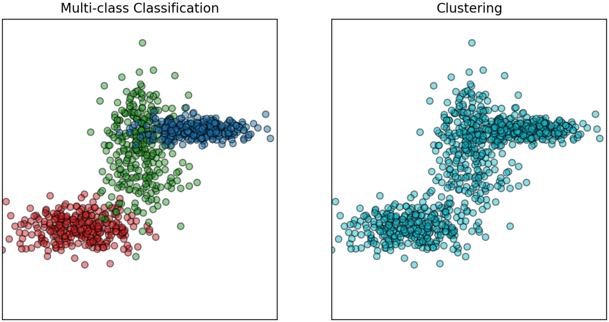
Prof. Reza Azadeh

University of Massachusetts Lowell

Clustering

- In this chapter, we discuss **clustering** methods.
- We consider the problem of identifying groups, or clusters, of data points in a multi-dimensional space.
- Suppose we have a data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ with N observations of a random D-dimensional Euclidean variable $\mathbf{x}$.
- Our goal is to partition the data set into some number of K of clusters.

Clustering



K-means Clustering (1)

- For each data point, $\mathbf{x}_n$, we introduce a corresponding set of binary indicator variables $r_{nk} \in \{0, 1\}$, where $k = 1, \ldots, K$ describing which of the K clusters the data point $\mathbf{x}_n$ is assigned to.
- If $\mathbf{x}_n$ is assigned to cluster k, we have $r_{nk} = 1$ and $r_{nj} = 0$ for $\forall j \neq k$.
- We define the following objective function known as *distortion measure*

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

where vector $\boldsymbol{\mu}_k$ represents represents the center of the cluster.

K-means Clustering (2)

- We aim to find the $\{r_{nk}\}$ and the $\{\boldsymbol{\mu}_k\}$ so as to minimize J.
- This can be done through an iterative process with each iteration including two successive optimization with respect to the r_{nk} and the $\boldsymbol{\mu}_k$.
  1. initialize $\boldsymbol{\mu}_k$
  2. minimize J w.r.t r_{nk}, keeping $\boldsymbol{\mu}_k$ fixed (*expectation*)
  3. minimize J w.r.t $\boldsymbol{\mu}_k$, keeping r_{nk} fixed (*maximization*)
  4. repeat until convergence

K-means Clustering (3) - Expectation Step

- The E-step, minimizing J with respect to r_{nk} has the following closed-form solution

$$r_{nk} = \begin{cases} 1 & \text{if } k = \operatorname{argmin}_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

K-means Clustering (4) - Maximization Step

- The M-step, minimize J with respect to $\boldsymbol{\mu}_k$, has the following solution

$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

K-means Clustering - example

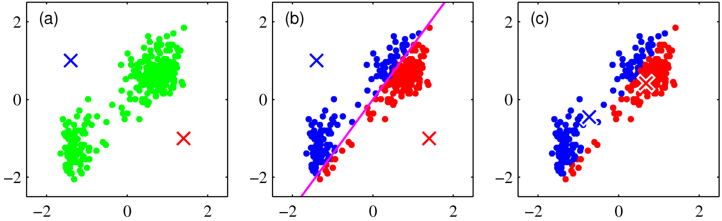


Figure: (a) data set in green, crosses show the initial choices for centers $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. (b) in the initial E-step, each data point is assigned either to the red or blue clusters, according to which cluster center is nearer. (c) in the subsequent M-step, each cluster center is recalculated to be the mean of the points assigned to the corresponding clusters.

K-means Clustering - example

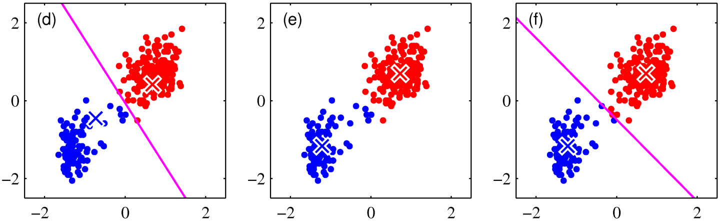


Figure: (e)-(f) more E-steps and M steps.

K-means Clustering - example

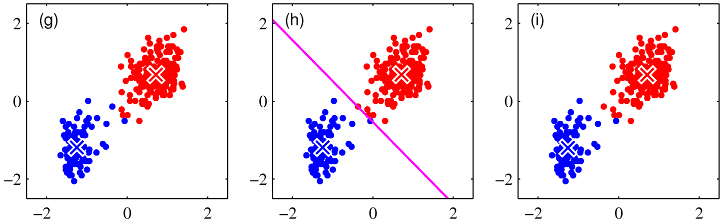


Figure: (g)-(i) successive E-steps and M steps through the final convergence of the algorithm.

K-means Clustering - example

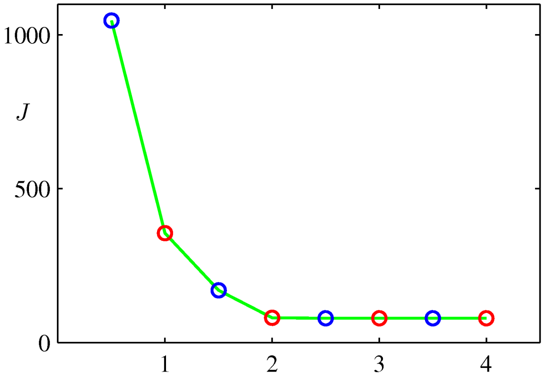


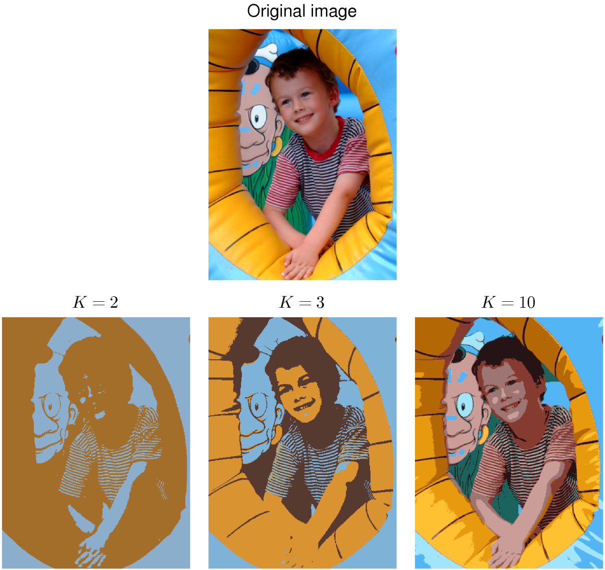
Figure: cost function J after each E-step (blue points) and M-step (red points) of the K-means algorithm. The algorithm has converged after the third M-step, and the final EM cycle produces no changes.

On-line K-means Clustering

- The E-step can be improved by using tree structures.
- We discussed the batch version of the algorithm, we can also drive an on-line stochastic algorithm by applying the Robinson-Monro procedure to the problem of finding the roots of the regression function. This leads to sequential update, in which for each data point $\mathbf{x}_n$ in turn, we update the nearest $\boldsymbol{\mu}_k$ using

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \eta_n(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{old}})$$

K-means Clustering for Image Segmentation



Mixture of Gaussians (1)

The Gaussian mixture distribution can be defined as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}_k).$$

We then introduce a K-dimensional binary random variable $\mathbf{z}$ having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are 0. The value of z_k therefore satisfies $z_k \in \{0, 1\}$ and $\sum_k z_k = 1$.

Mixture of Gaussians (2)

We shall define the joint distribution $p(\mathbf{x}, \mathbf{z})$ in terms of a marginal distribution $p(\mathbf{z})$ and a conditional distribution $p(\mathbf{x}|\mathbf{z})$. If we define $p(z_k = 1) = \pi_k$ where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$, the marginal can be written as

$$p(\mathbf{z}) = \Pi_{k=1}^{K} \pi_k^{z_k}$$

Mixture of Gaussians (3)

Similarly, we consider the conditional distribution of $\mathbf{x}$ given specific value for $\mathbf{z}$ is a Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{x}|\mathbf{z}) = \Pi_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

Mixture of Gaussians (4)

The marginal distribution of $p(\mathbf{x})$then can be written as

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Note that for every point $\mathbf{x}_n$ there is a corresponding latent variable $\mathbf{z}_n$.
Therefore, we have found and equivalent formulation for the Gaussian mixture involving an explicit latent variable which will lead to significant simplifications, through the introduction of Expectation-Maximization (EM) algorithm.

Mixture of Gaussians (5)

Another quantity that will play an important role in the conditional distribution $p(\mathbf{z}|\mathbf{x})$ which can be calculated using the Bayes' theorem

$$\begin{aligned}\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{k=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}\end{aligned}$$

where π_k is the prior probability of $z_k = 1$ and $\gamma(z_k)$ is the corresponding posterior probability once we have observed $\mathbf{x}$.

Mixture of Gaussians (6)

Suppose we have a data set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and we wish to model the data using a mixture of Gaussians.
We can represent the data set as an $N \times D$ matrix $\mathbf{X}$ in which the nth row is given by $\mathbf{x}_n^\top$. Similarly, the corresponding latent variables will be denoted by an $N \times K$ matrix $\mathbf{Z}$ with rows $\mathbf{z}_n^\top$.
By assuming points have been drawn independently from the distribution, we can write the log likelihood as

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \Big\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big\}$$

Mixture of Gaussians (7)

This log-likelihood function cannot be maximized in closed-form and should be solved through iterative optimization methods (e.g., gradient descent).

Here we use **Expectation-Maximization** which is a powerful method for finding maximum likelihood solutions for models with latent variables

EM for Gaussian Mixtures (1)

Given a Gaussian mixture model, the goal is to maximize the likelihood function w.r.t the parameters.

1. Initialize means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log-likelihood.
2. **E-step:** Evaluate the posterior using the current parameters

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

EM for Gaussian Mixtures (2)

3. **M-step:** Re-estimate the parameters using the current posterior

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N}\sum_{n=1}^{N}\gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^\top$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N}\gamma(z_{nk})$.

EM for Gaussian Mixtures (3)

4. Evaluate the log-likelihood and check for convergence of either the parameters or the log-likelihood. If the convergence conditions not satisfied return to step 2.

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

EM algorithm example

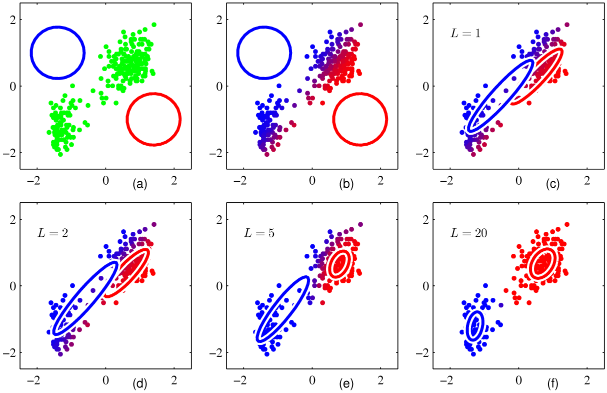


Figure: A mixture of two Gaussians applied to the 2D data set.

Notes

- The EM algorithm takes many more iterations to reach convergence compared with the K-means algorithm, and that each cycle requires significantly more computation.
- Therefore, it is common to run the K-means algorithm to find a suitable initialization for a Gaussian mixture model.
- Generally there will be multiple local maxima of the log likelihood and that EM algorithm is not guaranteed to find the global maximum.