Introduction to Big Data and Machine Learning
Mixture Models

Dr. Mihail

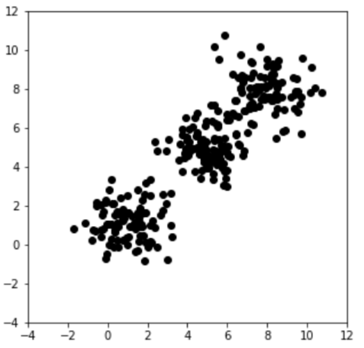
October 22, 2019

Mixture Models

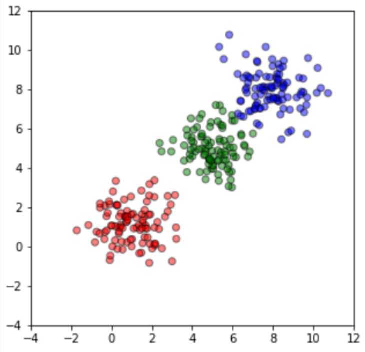
Idea

- Data can be made of complicated distributions that are formed from simpler compoenents.

Raw (unlabled)



Labeled



Mixture Models

In this lecture

- Consider the problem of finding clusters in a set of data points
- Non-probabilistic method, hard assignment
- Assumption is that the number of clusters is known
- K-Means algorithm
- A version of expectation maximization (EM) algorithm

Mixture Models

Problem Definition

- Consider the problem of identifying groups, or clusters, of data points in a multidimensional space
- Suppose the data set consists of $\{x_1, \ldots, x_N\}$, consisting of N D-dimensional observations
- Our goal is to partition the data into a set of K, where we will assume the number K is given (there are, however, ways to estimate it)

Mixture Models

Intuition

- A cluster, or group, is a set of data points where the inter-point distances within the group are small compared to the distances with points outside the cluster
- We formalize this by introducing a set of D-dimensional vectors μ_k, where $k = 1 \ldots k$, where μ_k is a prototype associated with the k^{th} cluster.
- μ_k can be thought of as the cluster centers

Goal

- Find cluster centers μ_k as well as an assignment of the data points to clusters, such that the sum of squares of the distances of each data point to its closest vector μ_k is a minimum.

Mixture Models

Notation

- For each data point 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 x_n is assigned to, so that if data point x_n is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$
- This is known as 1-of-K coding scheme

Mixture Models

Notation

- For each data point x_n, we introduce a corresponding set of binary indicator variables $r_{nk} \in \{0, 1\}$, where $k = 1, \dots, K$ describing which of the K clusters the data point x_n is assigned to, so that if data point x_n is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$
- This is known as 1-of-K coding scheme

Optimization problem

- We can then define an objective function:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2 \tag{1}$$

Mixture Models

Solving the optimization problem

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

- Goal is to find values for $\{r_{nk}\}$ and the $\{\mu_k\}$ so as to minimize J

Mixture Models

Solving the optimization problem

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||x_n - \mu_k||^2$$

- Goal is to find values for $\{r_{nk}\}$ and the $\{\mu_k\}$ so as to minimize J

Algorithm

First, choose initial values μ_k. Then, iterate two steps (EM):

1. Minimize J with respect to r_{nk}, keeping μ_k fixed (Expectation)
2. Minimize J with respect to μ_k, keeping r_{nk} fixed (Maximization)

Mixture Models

First step

- Consider determination of r_{nk}
- Since J is a linear function of r_{nk}, it has a closed-form solution
- We have n independent terms, each can be found in linear time, choose $r_{nk} = 1$ for whichever value of k gives the minimum value of $||x_n - \mu_k||^2$
- More formally:

$$r_{nk} = \begin{cases} 1 & if \;\; k = argmin_j ||x_n - \mu_j||^2 \\ 0 & otherwise \end{cases} \tag{2}$$

Mixture Models

Second step

- No consider optimizing μ_k, with r_{nk} fixed
- J is a quadratic function of μ_k, and it can be minimized by setting its derivative with respect to μ_k to zero, giving

$$2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0 \tag{3}$$

solving for μ_k gives:

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \tag{4}$$

Mixture Models

EM algorithm

- The steps above are repeated until no change in assignment is seen, or after a fixed number of iterations