Clustering, Mixture Models, and the EM Algorithm

Rebecca C. Steorts

Agenda

- Clustering
- Two Component Mixture Model
- Latent Variable
- EM Algorithm

Clustering

Clustering is an **unsupervised method** that divides up data into groups (clusters), so that points in any one group are more "similar" to each other than to points outside the group

Clustering methods (that we have covered)

- Fuzzy clustering (such as deterministic entity resolution or blocking)
- Overlapping clustering (such as locality sensitive hashing)
- Fellegi Sunter method (technically a mixture model, stay tuned)

Clustering methods (that we have not covered)

- Mixture models
- K-means
- Hierarchical clustering
- Others

Application areas

- Clustering temperatures to identify weather patterns, grouping individuals based on height and weight similarities.
- Clustering customers based on satisfaction levels (ordinal) or grouping individuals based on gender (categorical).
- Clustering GPS coordinates to identify spatial patterns, grouping locations on a map based on features.
- Clustering users in a social network or data based based on their connections (edge structure), grouping academic papers based on citation patterns.

History

- First proposed by Karl Pearson (1984) and analyzed on crab data.
- Applications: "agriculture, astronomy, bioinformatics, biology, economics, engineering, genetics, imaging, marketing, medicine, neuroscience, psychiatry, and psychology, among many other fields in the biological, physical, and social sciences". McLachlan et. al (2019).
- One of the methods in machine learning is **topic modeling**, which identifies "topics" in collections of documents/webpages.
- Topic modeling relies on mixtures models.

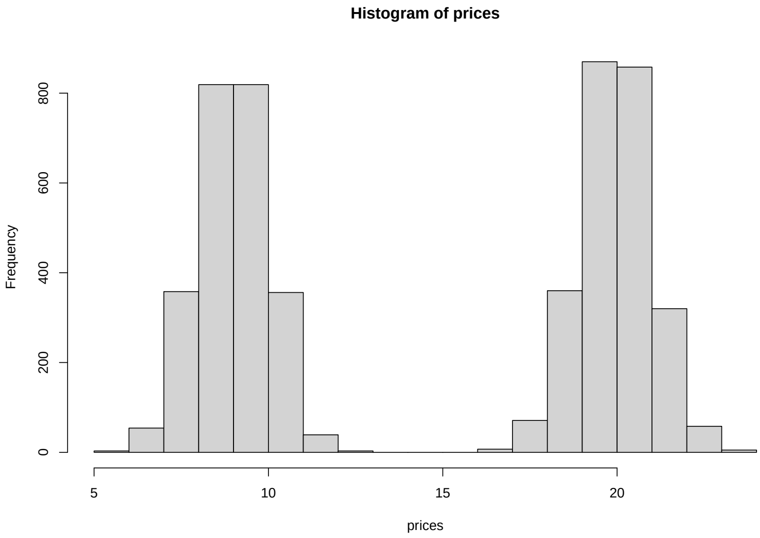
Motivation

- Suppose we want to simulate the price of a randomly chosen book.
- Paperbacks are often cheaper than hardbacks, so let's model them separately.
- Model the price of a book as a mixture model.
- There will be two components (or clusters) in our model – one for paperbacks and one for hardbacks.

Model

- Paperback distribution: $N(9, 1)$
- Hardback distribution: $N(20, 1)$
- Assume that there's a there is a 50% chance of choosing a paperback and 50% of choosing hardback.

Motivation

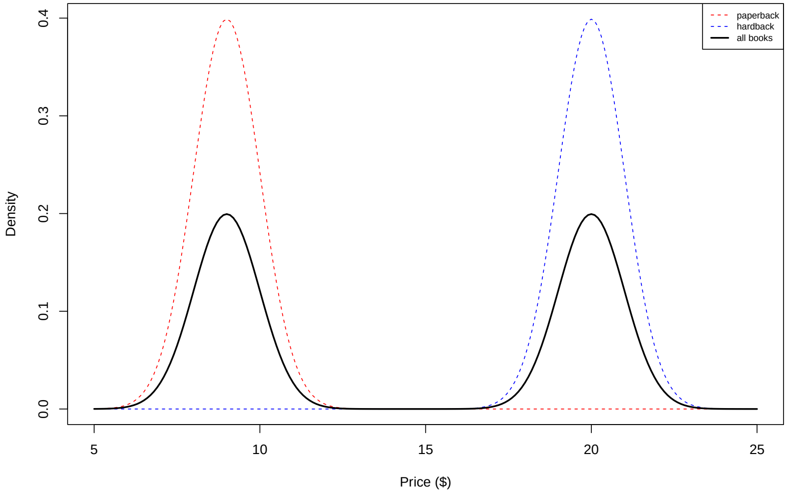


Histogram of prices
Frequency
0
200
400
600
800
5
10
15
20
prices

Motivation

- Are the prices of books unimodal or bimodal?
- Suppose you would want to predict the price of a book. Would its distribution be Normal or something else based on the the histogram.

Motivation



paperback
hardback
all books
Density
0.0
0.1
0.2
0.3
0.4
5
10
15
20
25
Price ($)

Motivation

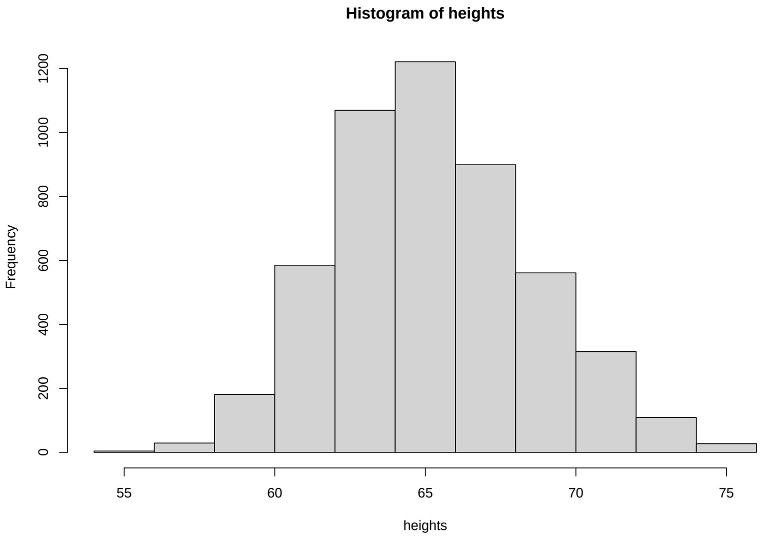
Now assume our data are the heights of students at university.

Male height: $N(69, 2.5^2)$, with units in inches.

Female height: $N(64, 2.5^2)$.

Assume that 75% of the population is female and 25% is male.

Motivation



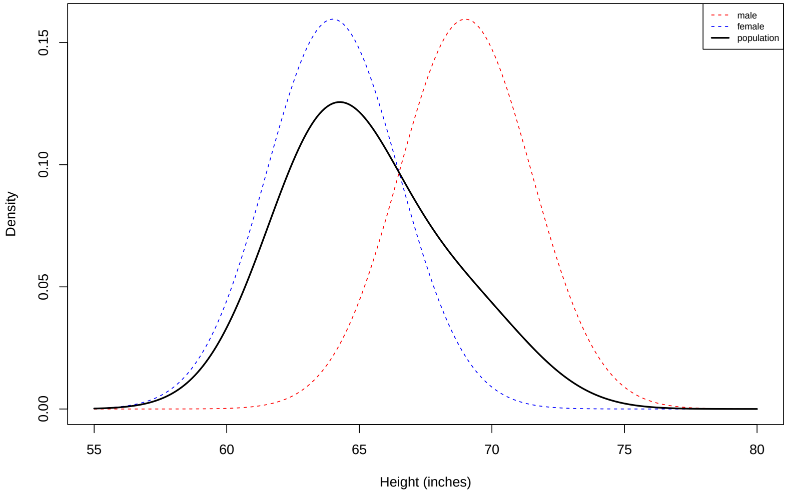
Histogram of heights
Frequency
0
200
400
600
800
1000
1200
55
60
65
70
75
heights

Motivation

The histogram is now unimodal.

Are heights normally distributed (assuming this model)? Let's invesitgate!

Motivation


male
female
population
0.00
0.05
0.10
0.15
Density
55
60
65
70
75
80
Height (inches)

Motivation

The Gaussian mixture model is unimodal because there is so much overlap between the two densities.

In this example, observe that the population density is not symmetric, and therefore not normally distributed.

Goal

The goal of this module is to introduce **mixture models**, which are commonly used in applications in classical and modern machine learning.

Mixture models can be viewed as probabilistic clustering

- Mixture models put similar data points into "clusters".
- This is appealing as we can potentially compare different probabilistic clustering methods by how well they predict (under cross-validation). We will not explore this in this particular lecture.
- This contrasts other methods such as k-means and hierarchical clustering as they produce clusters (and not predictions), so it's difficult to test if they are correct/incorrect.[1]

[1] Explore looking at these on your own and see if you can determine their limitations practically, compared to other machine learning models.

Two-component mixture model

Assume that both mixture components have the same precision, $\lambda = 1/\sigma^2$, which is fixed and known.

Let π be the mixture proportion for the first component.

Then the two-component Normal mixture model is:

$$X_1, \ldots, X_n \mid \mu, \pi \ \sim F(\mu, \pi) \tag{1}$$

where $F(\mu, \pi)$ is the distribution with p.d.f.

$$f(x|\mu, \pi) = (1 - \pi)\mathcal{N}(x \mid \mu_0, \lambda^{-1}) + \pi\mathcal{N}(x \mid \mu_1, \lambda^{-1}).$$

Likelihood

The likelihood is

$$
\begin{aligned}
p(x_{1:n}|\mu,\pi) &= \prod_{i=1}^{n} f(x_i|\mu,\pi) \\
&= \prod_{i=1}^{n} \Big[(1-\pi)\mathcal{N}(x_i \mid \mu_0,\lambda^{-1}) + \pi\mathcal{N}(x_i \mid \mu_1,\lambda^{-1})\Big].
\end{aligned}
$$

Likelihood

What do you notice about the likelihood function?

$$
\begin{aligned}
p(x_{1:n}|\mu,\pi) &= \prod_{i=1}^{n} f(x_i|\mu,\pi) \\
&= \prod_{i=1}^{n} \Big[(1-\pi)\mathcal{N}(x_i \mid \mu_0, \lambda^{-1}) + \pi\mathcal{N}(x_i \mid \mu_1, \lambda^{-1})\Big].
\end{aligned}
$$

Likelihood

The **likelihood** is very complicated function of μ and π.

This makes working with it directly to find the MLE (or other estimates) difficult.

Thus, we will rewrite the likelihood using a two-stage approach.

Two-stage approach

Let Z_i indicate whether subject i is from component 1 or 2.

$$Z_1, \ldots, Z_n \overset{iid}{\sim} \text{Bernoulli}(\pi) \tag{2}$$

$$X_i \mid Z \sim \mathcal{N}(\mu_{Z_i}, \lambda^{-1}) \quad i = 1, \ldots, n. \tag{3}$$

Checking for understanding

Then the two-component Normal mixture model is:

$$X_1, \ldots, X_n \mid \mu, \pi \;\sim F(\mu, \pi) \tag{4}$$

where $F(\mu, \pi)$ is the distribution with p.d.f.

$$f(x|\mu, \pi) = (1 - \pi)\mathcal{N}(x \mid \mu_0, \lambda^{-1}) + \pi\mathcal{N}(x \mid \mu_1, \lambda^{-1}).$$

Written as a two-stage process:

$$Z_1, \ldots, Z_n \mid \mu, \pi \overset{iid}{\sim} \text{Bernoulli}(\pi) \tag{5}$$

$$X_i \mid \mu, Z \sim \mathcal{N}(\mu_{Z_i}, \lambda^{-1}) i = 1, \ldots, n. \tag{6}$$

Checking for understanding

Given the two equivalent models above, how would you simulate data from a two component mixture model?

Extension to k-components

Assume we observe $X_1, \ldots, X_n$ and that each X_i is sampled from one of K **mixture components**.

Associated with each random variable X_i is a label called $Z_i \in \{1, \ldots, K\}$ which indicates which component X_i came from.

Notation

Let π_k be called **mixture proportions** or **mixture weights**, which represent the probability that X_i belongs to the k-th mixture component.

The mixture proportions are non-negative and they sum to one, $\sum_{k=1}^{K} \pi_k = 1$.

Observe that $P(X_i \mid Z_i = k)$ represents the distribution of X_i assuming it came from component k.

Extension

Then the two-component Normal mixture model is:

$$X_1, \ldots, X_n \mid \mu, \pi \ \sim F(\mu, \pi) \tag{7}$$

where $F(\mu, \pi)$ is the distribution with p.d.f.

$$f(x|\mu, \pi) = \sum_{k=1}^{K} \pi_k N(\mu_k, \lambda^{-1}).$$

Written as a two-stage process: for $i = 1, \ldots, n$:

$$P(Z_i = k) = \pi_k \tag{8}$$

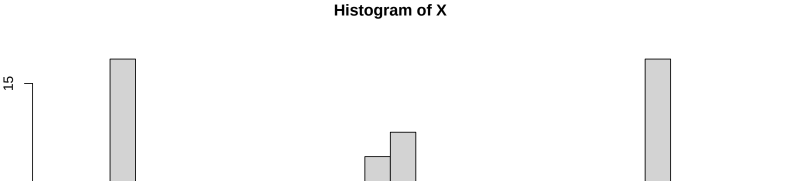
$$X_i \mid \mu, Z_i \sim \mathcal{N}(\mu_{Z_i}, \lambda^{-1}) \tag{9}$$

Example

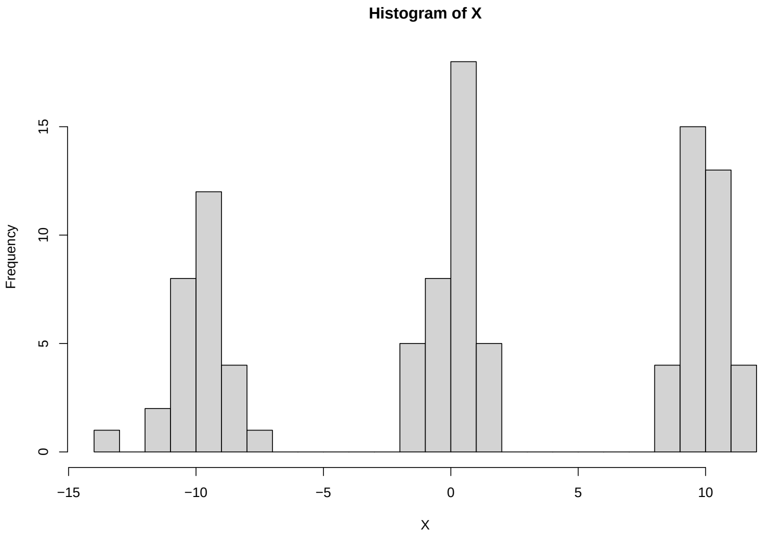
Let's look at a three component mixture model.

Suppose we assume that $\mu = (-10, 0, 10)$ and $\sigma^2 = 1$. Assume each mixture weight is equally likely.

```
set.seed(1234)
n <- 100
mu <- c(-10, 0, 10)
# sample Z first
Z <- sample(1:3, size=n, replace=TRUE)
# conditional on Z, sample the normal update
X <- rnorm(n, mean=mu[Z], sd=1)
hist(X, breaks=20)
```



Example



Estimation

Now assume we are in the Gaussian mixture model setting where the k-th component is $N(\mu_k, \sigma^2)$ and the mixture proportions are π_k.

How can we estimate $\{\mu_k, \sigma^2, \pi_k\}$ from the observed data $X_1, \ldots, X_n$?

Solution: EM Algorithm.

Conditional and marginal distributions

Recall that the conditional distribution $X_i | Z_i = k \sim N(\mu_k, \sigma_k^2)$, where $\pi_k = P(Z_i = k)$.

The marginal distribution of X_i is:

$$
\begin{aligned}
P(X_i = x) &= \sum_{k=1}^{K} P(Z_i = k) P(X_i = x | Z_i = k) && (10) \\
&= \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \sigma_k^2) && (11)
\end{aligned}
$$

Note: $\sigma_k^2 = \sigma^2$ moving forward.

Joint distribution

The joint probability of observations $X_1, \ldots, X_n$ is

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(x_i \mid \mu_k, \sigma_k^2)$$

Exercise

Show that

$$\log P(X_1, \ldots, X_n \mid \mu_1, \ldots, \mu_K) \tag{12}$$

$$= \log \prod_{i=1}^{n} P(x_i \mid \mu_1, \ldots, \mu_K) \tag{13}$$

$$= \sum_{i=1}^{n} \log[\sum_{k=1}^{K} P(x_i \mid \pi_k, \mu_1, \ldots, \mu_K)\pi_k] \tag{14}$$

Background

Recall that

$$\frac{\partial \log f(x)}{\partial dx} = \frac{1}{f(x)} \frac{\partial f(x)}{\partial dx}.$$

Exercise

Show that

$$\frac{\partial \log P(X_1, \ldots, X_n \mid \mu_1, \ldots, \mu_K)}{\partial \mu_k} \tag{15}$$

$$= \sum_{i=1}^{n} P(\pi_k \mid x_i, \mu_1, \ldots, \mu_K) \frac{(x_i - \mu_k)}{\sigma} \tag{16}$$

This implies that

$$\mu_k = \frac{\sum_{i=1}^{n} P(\pi_k \mid x_i, \mu_1, \ldots, \mu_K) x_i}{\sum_{i=1}^{n} P(\pi_k \mid x_i, \mu_1, \ldots, \mu_K)},$$

which is a non-linear equation of the μ_k's.

Intuition of EM

$$\mu_k = \frac{\sum_{i=1}^{n} P(\pi_k \mid x_i, \mu_1, \ldots, \mu_K) x_i}{\sum_{i=1}^{n} P(\pi_k \mid x_i, \mu_1, \ldots, \mu_K)},$$

- E-step: If for each x_i we knew that for each π_k the prob. that μ_k was in component π_k is $P(\pi_k \mid x_i, \mu_1, \ldots, \mu_K)$. Then we could compute μ_k.
- M-step: If we knew each μ_k, then we could compute $P(\pi_k \mid x_i, \mu_1, \ldots, \mu_K)$ for each μ_k and x_i

EM Algorithm

Initalize all the unknown parameters. On iteration t, let the estimates be $\lambda^{(t)} = \{\mu_1^{(t)}, \ldots, \mu_k^{(t)}\}$

1. E-Step:

$$P(\pi_k \mid x_i, \lambda^{(t)}) = \frac{P(\pi_k \mid x_i, \lambda^{(t)}) x_i}{P(\pi_k \mid x_i, \lambda^{(t)})} \tag{17}$$

2. M-Step:

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} P(\pi_k \mid x_i, \lambda^{(t)}) x_i}{\sum_{i=1}^{n} P(\pi_k \mid x_i, \lambda^{(t)})} \tag{18}$$

Exercise

Assume our mixture components are fully specified Gaussian distributions with $K = 2$ components:

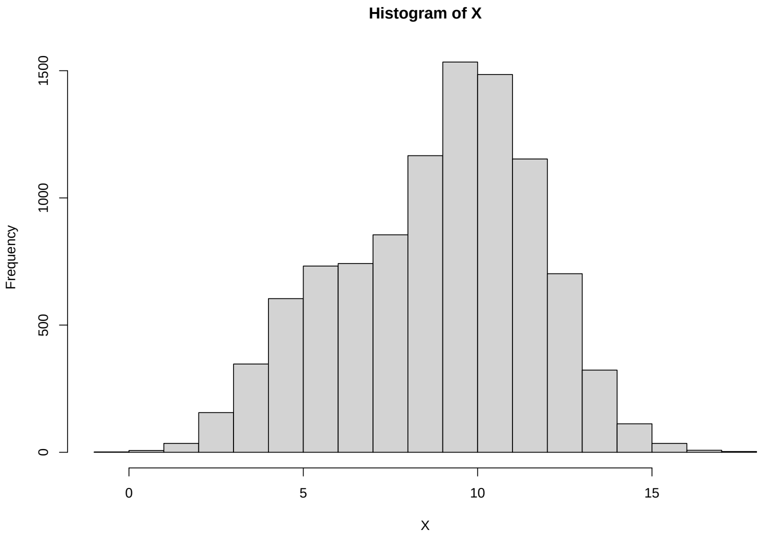
$$X_i \mid Z_i = 0 \sim N(5, 1.5) \tag{19}$$

$$X_i \mid Z_i = 1 \sim N(10, 2) \tag{20}$$

Let the true mixture proportions be $P(Z_i = 0) = 0.25$ and $P(Z_i = 1) = 0.75$, respectively.

Exercise

1. Simulate data from the mixture model on the previous slide, which should produce the following histogram.



Exercise

Compute the likelihood $P(X_i|Z_i = 0)$ and $P(X_i|Z_i = 1)$ and store these in a matrix L.

Exercise

Implement the E and M step in a function called `emIteration`. Then evaluate the EM and verify that your estimates are 0.29 and 0.71, respectively.

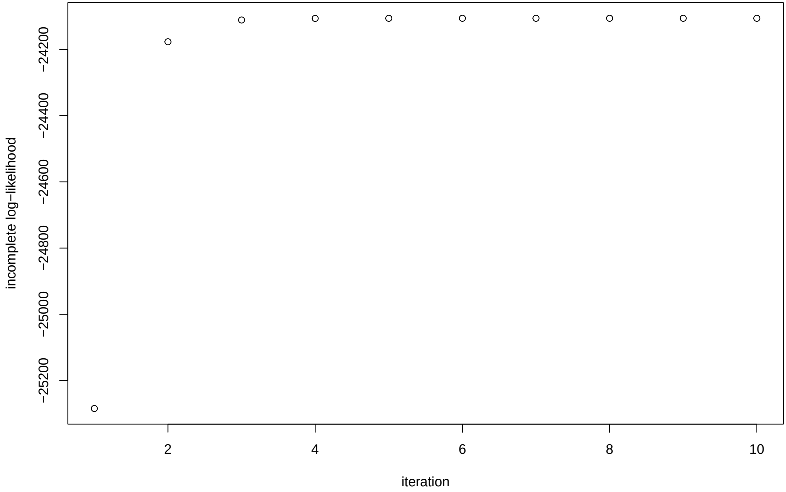
Plot the incomplete log-likelihood versus the iteration. What do you observe regarding its behavior.

Perform EM

```r
#perform EM
ee <- mixture.EM(w.init=c(0.5,0.5), L)
print(paste("Estimate = (", round(ee[[1]][1],2), ",",
            round(ee[[1]][2],2), ")", sep=""))
```

```
## [1] "Estimate = (0.23,0.77)"
```

Plot



The log-likelihood is strictly increasing, meaning that we have reached a local maxima.

R packages for mixture models

- The `mclust` package (http://www.stat.washington.edu/mclust/) is standard for Gaussian mixtures.
- The `mixtools` considers classic parametric densities, mixtures of regressions, and some non-parametric mixtures.

Exercise

Suppose that $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ independently.

1. What is the distribution of $aX + bY$?

Solution: Due to independence,

$$Z \sim N(a\mu_1, +b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2).$$

2. Suppose that $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ (and the observations are dependent).

Is the distribution of $aX + bY$ still Normal? No, not necessarily due to the dependence of the random variables.[2]

---

[2] In the case of a Gaussian mixture model, a random variable sampled from a Gaussian mixture model can be thought of as a two stage process. First, randomly sample a component (e.g., male or female). Second, then we sample our observation from the normal distribution that corresponds to the component sampled in step one.