Exam Review II and III Topics

Rebecca C. Steorts, Duke University

STA 325

Review of Exam II and III Material

Below I will review Exam II and III Material.

What is 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.

When might we want to use clustering?

One practical application of clustering is recommender systems, where one clusters users with similar viewing patterns on Netflix/Hulu, etc.

What are other applications we have seen in class or you have encountered?

Machine Learning Algorithms for Clustering

- k-means
- hierarchical clustering
- how to choose the number of clusters
- Mixture Models and the EM Algorithm

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).
- 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.

Mixture Model

Consider $X_1, \ldots, X_n$ 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.

Gaussian Mixture Model

Then the k-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.

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

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

$$P(Z_i = k) = \pi_k \tag{2}$$

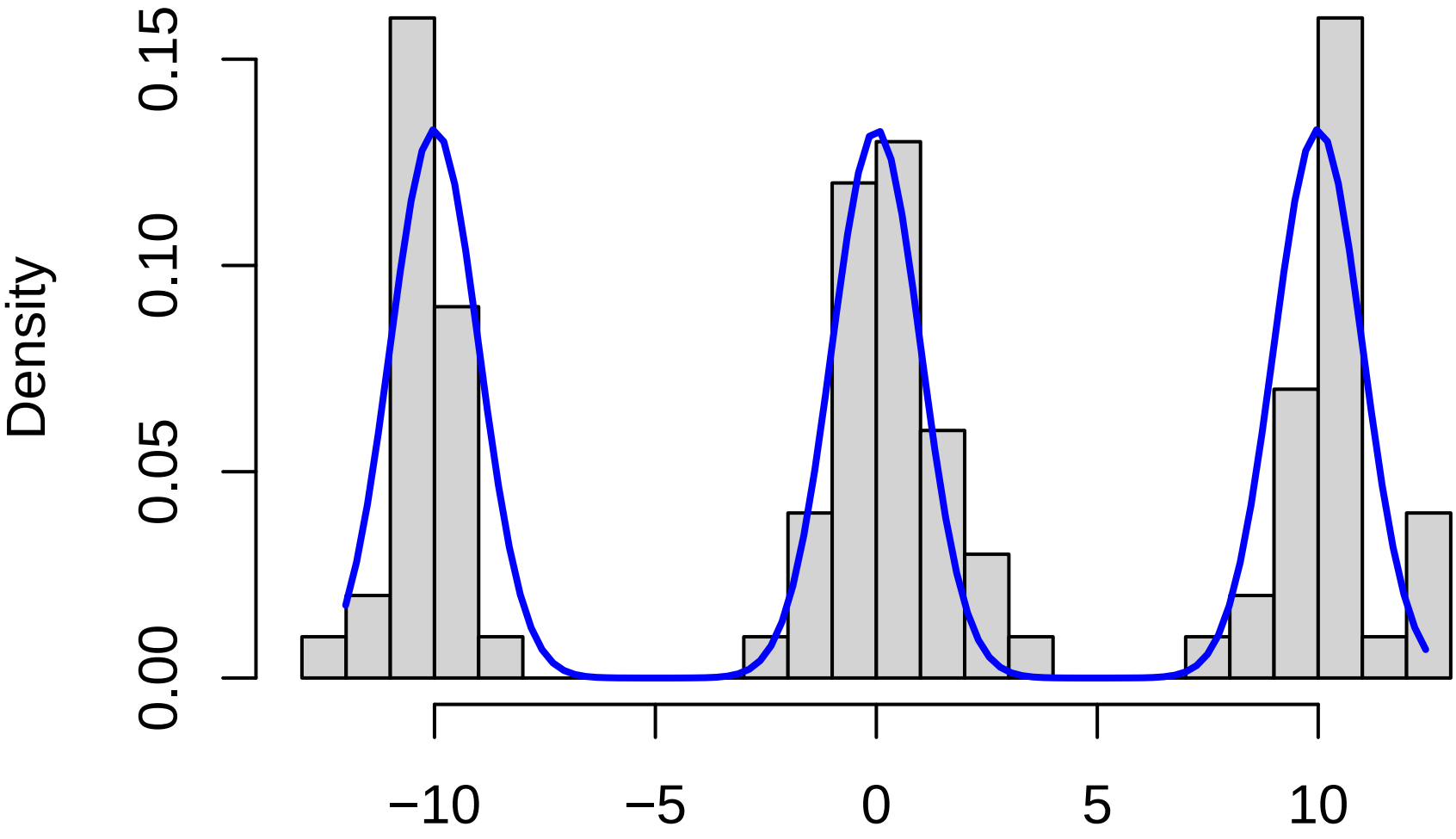
$$X_i \mid \mu, Z_i \sim \mathcal{N}(\mu_{Z_i}, \lambda^{-1}) \tag{3}$$

Can you Simulate Data from a Mixture Model?

Let's assume a three component mixture model.

Suppose we assume that $\mu = (-10, 0, 10)$ and $\sigma^2 = 1$. Assume each mixture weight is equally likely. How would you simulate data from this distribution and visualize it? Assume 100 data points.

Visualize the mixture model



How do we estimate the unknown parameters?

We use the Expectation-Maximation (EM) algorithm.

EM Algorithm

General way to deal with hidden class labels or clusters that are unknown such that we can estimate any unknown parameter values.

The method is extremely general and in fact and for more details on this please read through this tutorial on mixture models, the EM algorithm, and examples on Gaussian mixture models and Poisson mixture models.

EM Algorithm

The E stands for "expectation", because it gives us the conditional probabilities of different values of Z, and probabilities are expectations of indicator functions.

The M stands for "maximization."

The algorithm always converges to a local optima of the likelihood but not a global one.

EM Algorithm

Other resources:

- Tutorial on the Mixture Models and the EM algorithm: https://arxiv.org/pdf/1901.06708

The tutorial is quite general and has examples on both Gaussian and Poisson mixture models.

- Posted exercises on Exponential Mixture models
- Case Study on Snoq. Falls Data Set, where we looked at Gaussian Mixture Models

"Simple" EM Algorithm

Notation and Setup

We know the following:

- Observations $x_{1:n}$.
- K total classes
- $P(Z_i = k) = \pi_k$ (for $i = 1, \ldots, K$)
- Common variance σ^2.

We do not know $\mu_1, \ldots, \mu_K$ and want to learn these.

This is a very unrealistic setting, however, it hopefully provides intuition regarding the algorithm itself (and the math is simplified).

EM Algorithm

$\propto$ will drop any constants (and I will make sure to include them back in later). Common trick in Bayesian statistics.

$$p(x_1, \ldots, x_n \mid \mu_1, \ldots, \mu_K) \tag{4}$$

$$= \prod_{i=1}^{n} p(x_i \mid \mu_1, \ldots, \mu_K) \text{ independent data} \tag{5}$$

$$= \prod_{i=1}^{n} \sum_{k=1}^{K} p(x_i, z_i = k \mid \mu_1, \ldots, \mu_K) \text{ marg. over labels} \tag{6}$$

$$= \prod_{i=1}^{n} \sum_{k=1}^{K} p(x_i \mid z_i = k, \mu_1, \ldots, \mu_K) p(z_i = k) \tag{7}$$

$$\propto \prod_{i=1}^{n} \sum_{k=1}^{K} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_k)^2)\pi_k \text{ dropped normal constants} \tag{8}$$

EM Algorithm

Let $\theta^{(t)} = (\mu_1^{(t)}, \ldots, \mu_k^{(t)})$ at some iteration t.

At iteration t consider the function:

$$Q(\theta^{(t)} \mid \theta^{(t-1)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k \mid x_i, \theta^{(t-1)}) \tag{9}$$

$$\times \log P(x_i, z_i = k \mid \theta^{(t-1)}) \tag{10}$$

E-step

$$P(z_i = k \mid x_i, \theta^{t-1}) \tag{11}$$

$$= P(z_i = k \mid x_i, \mu_1^{(t-1)}, \ldots, \mu_K^{(t-1)}) \tag{12}$$

$$\propto P(x_i \mid z_i = k, \mu_1^{(t-1)}, \ldots, \mu_K^{(t-1)}) P(z_i = k) \tag{13}$$

$$\propto \exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k \tag{14}$$

$$= \frac{\exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k}{\sum_{k=1}^{K} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k} \tag{15}$$

This is equivalent to assigning clusters to each data point in a soft-way (clusters can overlap).

M-step

Recall that in the E-step, we calculated
$R_{ik}^{(t-1)} = P(z_i = k \mid x_i, \theta^{(t-1)})$

$$Q(\theta^{(t)} \mid \theta^{(t-1)}) \tag{16}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k \mid x_i, \theta^{(t-1)}) \times \log P(x_i, z_i = k \mid \theta^{(t-1)}) \tag{17}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} P(z_i = k \mid x_i, \theta^{(t-1)}) \tag{18}$$

$$\times [\log P(x_i \mid z_i = k, \theta^{(t-1)}) + \log P(z_i = k \mid \theta^{(t-1)})] \tag{19}$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} R_{ik}^{(t-1)} [-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2) + \log \pi_k] \tag{20}$$

M-step

At each iteration t, maximize Q in term of $\theta^{(t)}$.

$$Q(\mu_k^{(t)} \mid \theta^{(t-1)}) \propto \sum_{i=1}^{n} R_{ik}^{(t-1)}(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2), \implies \quad (21)$$

$$\frac{\partial Q(\mu_k^{(t)} \mid \theta^{(t-1)})}{\partial \mu_k^{(t)}} = \sum_{i=1}^{n} R_{ik}^{(t-1)}(x_i - \mu_k^{(t-1)})) = 0 \implies \quad (22)$$

$$\mu_k^{(t)} = \sum_{i=1}^{n} w_i x_i \quad \text{where}$$

$$w_i = \frac{R_{ik}^{t-1}}{\sum_{i=1}^{n} R_{ik}^{t-1}} = \frac{P(z_i = k \mid x_i, \theta^{(t-1)})}{\sum_{i=1}^{n} P(z_i = k \mid x_i, \theta^{(t-1)})}$$

This is equivalent to updating the cluster centers.

Summarize EM Algorithm

1. E-step

Compute the expected classes of all data points for each class:

$$P(z_i = k \mid x_i, \theta^{(t-1)}) = \frac{\exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k^{(t-1)}}{\sum_{k=1}^{K} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_k^{(t-1)})^2)\pi_k^{(t-1)}}$$

2. M-step

Then compute the maximum value given our data's class membership.

$$\mu_i^{(t)} = \sum_{i=1}^{n} w_i x_i.$$

In this case, it's the MLE but with weighted data.