

GMM and EM questions

Haresh Rengaraj Rajamohan

CDS, NYU

April 28, 2021

Recap: Gaussian Mixture models

Question 1: Clustering¹

Consider the set of training data below, and two clustering algorithms: K-Means, and a Gaussian Mixture Model (GMM) trained using EM. Will these two clustering algorithms produce the same cluster centers (means) for this data set? In one sentence, explain why or why not



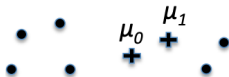
¹From CMU

[Solution] Question 1: Clustering Comparison

- Both the approaches will find the clusters
- In k-means the center of a cluster is the average of all the elements in the cluster
- In GMM, the centers are weighted average of all the elements in the data.
- So, in GMM, we can expect the right center to be skewed a bit to the left and left center to the right

Question 2: EM basics ²

Consider applying EM to train a GMM to cluster the data into two clusters. The '+' points indicate the current means μ_0 , μ_1 of the two components of the mixture after the k th iteration of EM.

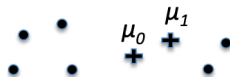


- Draw on the figure the directions in which μ_0 and μ_1 will move during the next M-step
- Will the marginal likelihood of the training data, increase or decrease on the next EM iteration?
- Will the estimate of π_0 increase or decrease on the next EM step?

²From CMU

[Solution] Question 2: EM basics

Consider applying EM to train a GMM to cluster the data into two clusters. The '+' points indicate the current means μ_0 , μ_1 of the two components of the mixture after the k th iteration of EM.



- μ_0 moves to the left, and μ_1 moves to the right.
- Increase. Each iteration of the EM algorithm increases the likelihood of the data, unless you happen to be exactly at a local optimum
- It will increase

Question 3: Gaussian Naive Bayes and GMMs ³

Lets consider the relationship between a Gaussian Naive Bayes (GNB) classifier and the above Gaussian Mixture Model (GMM). It is easy to see that they involve the same probabilistic model. Our usual GNB classifier assumes $p(Y|X)$ is of the form:

$$p(Y|X) = \frac{P(Y)\prod_i p(X_i|Y)}{p(X)}$$

where Y is a Bernoulli random variable (i.e., $P(Y = 0) = \pi_0$). It also assumes each feature X_i is governed by a Gaussian distribution conditioned on Y . For simplicity, let's assume all features have the same variance, so

$$P(X_i|Y = k) \sim N(\mu_{k,i}, \sigma)$$

³From CMU

Question 3: Gaussian Naive Bayes and GMMs continued

Notice this GNB generative model is identical to that of our GMM above (plus the simplifying assumption of identical σ s). In other words, both models assume we generate data points by choosing a Y according to π_0 , then drawing an X according to a Gaussian conditioned on Y .

- The GNB objective is $\operatorname{argmax}_{\theta} \prod_j P(x^j, y^j | \theta)$. Give the EM objective for GMM.
- Suppose we have a set of training data in which we have both labeled and unlabeled samples. We have known y values for x^1, \dots, x^m but have additional unlabeled examples x^{m+1}, \dots, x^{m+n} without known values for y . Propose a modified EM approach to train in this setting.
- Write down the objective function that your modified EM is maximizing. In your expression, distinguish between the labeled and unlabeled examples.

[Solution] Question 3: Gaussian Naive Bayes and GMMs

- $\operatorname{argmax}_{\theta} \Pi_j \left(\sum_y P(x^j, y | \theta) \right)$
- In the E step, for labeled samples use
- $\gamma_{ij} = \delta_{j, y(i)} \operatorname{argmax}_{\theta} \Pi_{j=1}^m \left(\sum_y P(x^j, y | \theta) \right) \Pi_{j=m+1}^{m+n} (P(x^j, y^j | \theta))$, where $\delta_{j, y(i)} = 1(j = y(i))$

Question 4: EM computation ⁴

Suppose that we are fitting a Gaussian mixture model for data items consisting of a single real value, x , using $K = 2$ components. We have $N = 5$ training cases, in which the values of x are as follows:

5, 15, 25, 30, 40

We use the EM algorithm to find the maximum likelihood estimates for the model parameters, which are the mixing proportions for the two components, π_1 and π_2 , and the means for the two components, μ_1 and μ_2 . The standard deviations for the two components are fixed at 10.

⁴From UToronto

Question 4: EM computation continued

Suppose that at some point in the EM algorithm, the E step found that the responsibilities of the two components for the five data items were as follows:

| r_{i1} | r_{i2} |
|----------|----------|
| 0.2 | 0.8 |
| 0.2 | 0.8 |
| 0.8 | 0.2 |
| 0.9 | 0.1 |
| 0.9 | 0.1 |

What values for the parameters π_1 , π_2 , μ_1 , and μ_2 will be found in the next M step of the algorithm?

[Solution] Question 4 EM computation

The new estimates will be

- $\pi_1 = (0.2 + 0.2 + 0.8 + 0.9 + 0.9)/5 = 0.6$
- $\pi_2 = (0.8 + 0.8 + 0.2 + 0.1 + 0.1)/5 = 0.4$
- $\mu_1 = (0.2 \times 5 + 0.2 \times 15 + 0.8 \times 25 + 0.9 \times 30 + 0.9 \times 40)/(0.2 + 0.2 + 0.8 + 0.9 + 0.9) = 29$
- $\mu_2 = (0.8 \times 5 + 0.8 \times 15 + 0.2 \times 25 + 0.1 \times 30 + 0.1 \times 40)/(0.8 + 0.8 + 0.2 + 0.1 + 0.1) = 14$

Question 5 Computation Problem ⁵

Consider a two-component Gaussian mixture model for univariate data, in which the probability density for an observation, x , is,

$$\frac{1}{2}N(x|\mu, 1) + \frac{1}{2}N(x|\mu, 2^2)$$

Here, $N(x|\mu, \sigma^2)$ denotes the density for x under a univariate normal distribution with mean μ and variance σ^2 . Notice that mixing proportions are equal for this mixture model, that the two components have the same mean, and that the standard deviations of the two components are fixed at 1 and 2. There is only one model parameter, μ .

⁵From UToronto

Question 5 Computation Problem continued

Suppose we wish to estimate the μ parameter by maximum likelihood using the EM algorithm. Answer the following questions regarding how the E step and M step of this algorithm operate, if we have the three data points below:

4.0, 4.6, 2.0

- Find the responsibilities that will be computed in the E step if the model parameter estimates from the previous M step are $\mu = 4$, $\sigma_1 = 1$, and $\sigma_2 = 2$. Since the responsibilities for the two components must add to one, it is enough to give $r_{i1} = P(\text{component}_1 | x_i)$ for $i = 1, 2, 3$.
- Using the responsibilities that you computed in part (a), find the estimate for μ that will be found in the next M step.

[Solution] Question 5 Computation Problem

(A). Using Bayes Rule,

$$P(\text{component 1} | x) = \frac{\frac{1}{2} N(x | \mu, 1)}{\frac{1}{2} N(x | \mu, 1) + \frac{1}{2} N(x | \mu, 2^2)}$$

Lets apply this to the three observations,

$$r_{11} = \frac{(1/2)0.4}{(1/2)0.4 + (1/2)(1/2)0.4} = 2/3$$

$$r_{21} = \frac{(1/2)0.33}{(1/2)0.33 + (1/2)(1/2)0.38} = 33/52$$

$$r_{31} = \frac{(1/2)0.05}{(1/2)0.05 + (1/2)(1/2)0.24} = 5/17$$

[Solution continued] Question 5 Computation Problem

(B). The expected log likelihood is,

$$\sum_{i=1}^3 [r_{i1}(-1/2)(x_i - \mu)^2 + (1 - r_{i1})(-1/2)(x_i - \mu)^2/4]$$

Lets differentiate and equate this to 0,

$$\sum_{i=1}^3 [r_{i1}(x_i - \mu) + (1 - r_{i1})(x_i - \mu)/4] = 0$$

We get,

$$\hat{\mu} = \frac{\sum_{i=1}^3 (r_{i1} + (1 - r_{i1})/4)x_i}{\sum_{i=1}^3 (r_{i1} + (1 - r_{i1})/4)} = \frac{(3/4)4.0 + (151/208)4.6 + (25/68)2.0}{(3/4) + (151/208) + (25/68)}$$

Question 6 EM derivation ⁶

Lets derive the E-M update rules for a univariate Gaussian mixture model (GMM) with two mixture components. Unlike the GMMs we covered in the course, the mean μ will be shared between the two mixture components, but each component will have its own standard deviation σ_k . The model is defined as follows:

$$z \sim \text{Bernoulli}(\theta)$$

$$x|z = k \sim N(\mu, \sigma_k)$$

- Write the density defined by this model (i.e. the probability of x , with z marginalized out)
- E-Step: Compute the posterior probability $r^{(i)} = P(z^{(i)} = 1|x^{(i)})$
- Update rule for μ (σ_k fixed) and σ_k (μ fixed)

⁶From UToronto

[Solution] Question 6 EM derivation

(A).

$$p(x) = \theta N(x; \mu, \sigma_1) + (1 - \theta) N(x; \mu, \sigma_0)$$

(B).

$$r^{(i)} = \frac{\theta N(x; \mu, \sigma_1)}{\theta N(x; \mu, \sigma_1) + (1 - \theta) N(x; \mu, \sigma_0)}$$

(C).

$$\mu \leftarrow \frac{\sum_{i=1}^N x^{(i)} (r^{(i)} \sigma_0^2 + (1 - r^{(i)}) \sigma_1^2)}{\sum_{i=1}^N (r^{(i)} \sigma_0^2 + (1 - r^{(i)}) \sigma_1^2)}$$

$$\sigma_1^2 \leftarrow \frac{\sum_{i=1}^N r^{(i)} (x^{(i)} - \mu)^2}{\sum_{i=1}^N r^{(i)}}$$

Question 7.1 MCQ ⁷

Assume you have points that are generated by one of two possible Gaussian distributions. Which of the following are true?

- We know how to get a globally optimal solution by deriving the maximum likelihood estimate analytically
- Using the EM algorithm to solve this problem, we assume that we know from which Gaussian each point originated.
- Once the EM algorithm has converged, we know for certain from which Gaussian each point originated.
- The EM algorithm for this problem guarantees that the likelihood of the data never decreases from one iteration to the next.

⁷from CMU

[Solution] Question 7.1 MCQ

Answer: (d). A - EM doesn't give the globally optimal solution. B - We can start out with one of the Gaussians being more likely for some points, but we don't know for sure. C - After convergence, we only know the probability values of belonging to a particular Gaussian.

Question 7.2 MCQ ⁸

Which of the following are true about the EM algorithm as applied to a Gaussian Mixture Model?

- The choice of initial values of parameters of the Gaussian affects the final estimates.
- The algorithm is guaranteed to converge
- The algorithm is guaranteed to converge to a global maxima.
- The estimate of the parameters obtained at the end is the Maximum Likelihood Estimate.

⁸from CMU

[Solution] Question 7.2 MCQ

A and B are true. C - EM doesn't give the globally optimal solution. D - We cannot solve GMM in closed form to get a clean maximum likelihood expression

- GMM tutorial