Course > Unit 4 Unsupervised Learning (2 weeks) > Lecture 16. Mixture Models; EM algorithm > 5. Mixture Model - Unobserved Case: EM Algorithm

5. Mixture Model - Unobserved Case: EM Algorithm

The EM Algorithm

Video

Download video file

Transcripts

Download SubRip (.srt) file
Download Text (.txt) file

Estimates of Parameters of GMM: The Expectation Maximization (EM) Algorithm

We observe n data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbb{R}^d$. We wish to maximize the GMM likelihood with respect to the parameter set $\theta = \left\{p_1, \ldots, p_K, \mu^{(1)}, \ldots, \mu^{(K)}, \sigma_1^2, \ldots, \sigma_K^2\right\}$.

Maximizing the log-likelihood $\log\left(\prod_{i=1}^n p\left(\mathbf{x}^{(i)}|\theta\right)\right)$ is not tractable in the setting of GMMs. There is no closed-form solution to finding the parameter set θ that maximizes the likelihood. The **EM algorithm** is an iterative algorithm that finds a locally optimal solution $\hat{\theta}$ to the GMM likelihood maximization problem.

E Step

The **E Step** of the algorithm involves finding the posterior probability that point $\mathbf{x}^{(i)}$ was generated by cluster j, for every $i = 1, \ldots, n$ and $j = 1, \ldots, K$. This step assumes the knowledge of the parameter set θ. We find the posterior using the following equation:

$$p\left(\text{point } \mathbf{x}^{(i)} \text{ was generated by cluster } j|\mathbf{x}^{(i)}, \theta\right) \triangleq p\left(j \mid i\right) = \frac{p_j \mathcal{N}\left(\mathbf{x}^{(i)}; \mu^{(j)}, \sigma_j^2 I\right)}{p\left(\mathbf{x}^{(i)} \mid \theta\right)}.$$

M Step

The **M Step** of the algorithm maximizes a proxy function $\hat{\ell}\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \mid \theta\right)$ of the log-likelihood over θ, where

$$\hat{\ell}\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \mid \theta\right) \triangleq \sum_{i=1}^{n} \sum_{j=1}^{K} p\left(j \mid i\right) \log\left(\frac{p\left(\mathbf{x}^{(i)} \text{ and } \mathbf{x}^{(i)} \text{ generated by cluster } j \mid \theta\right)}{p\left(j \mid i\right)}\right).$$

This is done instead of maximizing over θ the actual log-likelihood

$$\ell\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \mid \theta\right) = \sum_{i=1}^{n} \log\left[\sum_{j=1}^{K} p\left(\mathbf{x}^{(i)} \text{ generated by cluster } j \mid \theta\right)\right].$$

Maximizing the proxy function over the parameter set θ, one can verify by taking derivatives and setting them equal to zero that

$$\begin{aligned}\widehat{\mu^{(j)}} &= \frac{\sum_{i=1}^n p\left(j \mid i\right) \mathbf{x}^{(i)}}{\sum_{i=1}^n p\left(j \mid i\right)} \\ \widehat{p_j} &= \frac{1}{n} \sum_{i=1}^{n} p\left(j \mid i\right),\end{aligned}$$

$$\widehat{\sigma_j^2} = \frac{\sum_{i=1}^{n} p\left(j \mid i\right) \left\|\mathbf{x}^{(i)} - \widehat{\mu^{(j)}}\right\|^2}{d \sum_{i=1}^{n} p\left(j \mid i\right)}.$$

The E and M steps are repeated iteratively until there is no noticeable change in the actual likelihood computed after M step using the newly estimated parameters or if the parameters do not vary by much.

Initialization

As for the initialization before the first time E step is carried out, we can either do a random initialization of the parameter set θ or we can employ k-means to find the initial cluster centers of the K clusters and use the global variance of the dataset as the initial variance of all the K clusters. In the latter case, the mixture weights can be initialized to the proportion of data points in the clusters as found by the k-means algorithm.

Gaussian Mixture Model: An Example Update - E-Step

5/5 points (graded)

Assume that the initial means and variances of two clusters in a GMM are as follows: $\mu^{(1)} = -3$, $\mu^{(2)} = 2$, $\sigma_1^2 = \sigma_2^2 = 4$. Let $p_1 = p_2 = 0.5$.

Let $x^{(1)} = 0.2$, $x^{(2)} = -0.9$, $x^{(3)} = -1$, $x^{(4)} = 1.2$, $x^{(5)} = 1.8$ be five points that we wish to cluster.

In this problem and in the next, we compute the updated parameters corresponding to cluster 1. You may use any computational tool at your disposal.

Compute the following posterior probabilities (provide at least five decimal digits):

$p\left(1 \mid 1\right) =$

0.294215 ✓ **Answer:** 0.29421

$p\left(1 \mid 2\right) =$

0.62246 ✓ **Answer:** 0.62246

$p\left(1 \mid 3\right) =$

0.651355 ✓ **Answer:** 0.65135

$p\,(1 \mid 4) =$

0.10669 ✓ **Answer:** 0.10669

$p\,(1 \mid 5) =$

0.05340 ✓ **Answer:** 0.053403

Solution:

Using the formula of the E-step

$$p\,(j \mid i) = \frac{p_j\,\mathcal{N}\,(x^{(i)};\mu^{(j)},\sigma_j^2)}{p\,(x^{(i)}|\theta)},$$

we can obtain that

$$\begin{aligned} p\,(1 \mid 1) &= 0.29421 \\ p\,(1 \mid 2) &= 0.62246 \\ p\,(1 \mid 3) &= 0.65135 \\ p\,(1 \mid 4) &= 0.10669 \\ p\,(1 \mid 5) &= 0.053403. \end{aligned}$$

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

Gaussian Mixture Model: An Example Update - M-Step

2/3 points (graded)

Compute the updated parameters corresponding to cluster 1 (provide at least five decimal digits):

$\hat{p}_1 =$

0.34562 ✓ **Answer:** 0.34562

$\hat{\mu}_1 =$

-0.53733 ✓ **Answer:** -0.53733

$\hat{\sigma}_1^2 =$

5.42245 ✗ **Answer:** 0.57579

Solution:

Using the formulae corresponding to the M-step,

$$\begin{aligned}
\hat{n}_1 &= \sum_{i=1}^{5} p\left(1|i\right) = 1.7281 \\
\hat{p}_1 &= \frac{\hat{n}_1}{n} = \frac{\hat{n}_1}{5} = 0.34562 \\
\hat{\mu}_1 &= \frac{1}{\hat{n}_1} \sum_{i=1}^{5} p\left(1|i\right) x^{(i)} = -0.53733 \\
\hat{\sigma}_1^2 &= \frac{1}{\hat{n}_1} \sum_{i=1}^{5} p\left(1|i\right) \left(x^{(i)} - \hat{\mu}^{(1)}\right)^2 = 0.57579.
\end{aligned}$$

Submit You have used 3 of 3 attempts

Answers are displayed within the problem

Gaussian Mixture Model and the EM Algorithm

1/1 point (graded)

Which of the following statements are true? Assume that we have a Gaussian mixture model with known (or estimated) parameters (means and variances of the Gaussians and the mixture weights).

☑ A Gaussian mixture model can provide information about how likely it is that a given point belongs to each cluster.

□ The EM algorithm converges to the same estimate of the parameters irrespective of the initialized values.

□ An iteration of the EM algorithm is computationally more expensive (in terms of order complexity) when compared to an iteration of the K-means algorithm for the same number of clusters.

✓

Solution:

The first statement is true because the estimated posterior probabilities tell us how likely it is that a given point belongs to each cluster. The third statement is true because each iteration of the EM algorithm involves two steps that are each more computationally expensive than the updates involved in an iteration of the K-means algorithm.

The second statement is not true. As explained in the video, the EM algorithm is guaranteed (under some conditions) to only converge locally.

The third statement is also not true. We can see that the E-step of the algorithm takes $\mathcal{O}\left(nKd\right)$ computations and the M-step of the algorithm is also of the order $\mathcal{O}\left(nKd\right)$.

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Mixture Models and Digit Classification

3/3 points (graded)

Assume that we have 100,000 black-and-white images of size 26×26 pixels that are the result of scans of hand-written digits between 0 and 9.

We can apply mixture models to effectively train a classifier based on clustering using the EM algorithm applied to the dataset.

Identify the following parameters (according to notation developed in the lecture, assuming that we use all the data for training):

$K =$

10 ✓ **Answer:** 10

$n =$

100000 ✓ **Answer:** 100000

$d =$

676 ✓ **Answer:** 676

Solution:

We are classifying $n = 100,000$ vectors each of length $d = 676$ into $K = 10$ clusters (one cluster for each digit).

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Note: The Gaussian mixture model can be extended to the case where each mixture component has a general covariance matrix Σ_j. The case that we have studied so far is a special case where $\Sigma_j = \sigma_j^2 I$, where I is the identity matrix of size $d \times d$. The EM algorithm can also be extended to work in this general setting.

Discussion

Hide Discussion

Topic: Unit 4 Unsupervised Learning (2 weeks) :Lecture 16. Mixture Models; EM algorithm / 5. Mixture Model - Unobserved Case: EM Algorithm

Add a Post

Show all posts by recent activity

EM explanation

Pinned 5

? [Staff] Two contradicting explanations - Gaussian Mixture Model and the EM Algorithm

There are two contradicting explanations for the third choice in the answer. Please check. 5

- ☑ Notation for the normal distribution 4
  The question may be trivial but I'm not getting anywhere. The notation N(x, mu, sigma) stands for the density of t...

- 💬 Impact of the relative sizes of the sigmas 5
  👤 Community TA

- 💬 Wrong Instruction in E Step problem? 9
  I used up my three attempts to solve this problem, but seems like the instruction is wrong in the first place. It see...

- ☑ How to compute the following in python 5
  Based on the following, how can this be computed in python? N(x(i);μ(j),σ2jI)

- ? How would you do the EM-algorithm for points with more features? 3
  How would you compute f.e. p(j|i) ? by simply taking the norm of the vector x_(i) or would you take N(x_(i),...) for e...

- 💬 M step 2

- ☑ formula for update of sigma^2 value in M step 2
  Is the formula provided for updating the sigma^2 value in the M step correct?

- ☑ Need Simple Example - Gaussian Mixture Model: An Example Update - E-Step 2
  Looking for insight. I am getting the wrong answer by trying to follow all the right steps. I'm beginning to wonder i...