5. Mixture Model - Unobserved Case: EM Algorithm

Exercises due Apr 19, 2023 08:59 -03 Completed

The EM Algorithm

▶ 0:00 / 0:00 1.0x

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

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 $\log(\prod_{i=1}^n p(\mathbf{x}^{(i)} \mid \theta))$ is not tractable in the setting of GMMs. solution to finding the parameter set θ that maximizes the likelihood. The **EM 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 $\mathbf{x}^{(i)}$ wa for every $i = 1, \ldots, n$ and $j = 1, \ldots, K$. This step assumes the knowledge of the parameter

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

Maximizing the proxy function over the parameter set θ, one can verify by taking deriva equal to zero that

$$\hat{\mu}^{(j)} = \frac{\sum_{i=1}^{n} p(j \mid i)\mathbf{x}^{(i)}}{\sum_{i=1}^{n} p(j \mid i)}$$

$$\hat{p}_j = \frac{1}{n} \sum_{i=1}^{n} p(j \mid i),$$

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} p(j \mid i)\|\mathbf{x}^{(i)} - \hat{\mu}^{(j)}\|^2}{d \sum_{i=1}^{n} p(j \mid i)}.$$

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

Initialization

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

---

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: μ $\sigma_1^2 = \sigma_2^2 = 4$. 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 t

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

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

$p(1 \mid 1) =$

$p(1 \mid 5) =$

0.05340

Submit You have used 3 of 3 attempts

Gaussian Mixture Model: An Example Update - M-Step

2/3 points (graded)

Compute the updated parameters corresponding to cluster 1 (provide at least five deci

$\hat{p}_1 =$

0.34562

$\hat{\mu}_1 =$

-0.53733

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

0.05340

Submit You have used 3 of 3 attempts

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 m estimated) parameters (means and variances of the Gaussians and the mixture weights

- [x] A Gaussian mixture model can provide information about how likely it is that a giv each cluster.
- [ ] The EM algorithm converges to the same estimate of the parameters irrespective values.

to the dataset.

Identify the following parameters (according to notation developed in the lecture, assu
data for training):

$K =$

10

$n =$

100000

$d =$

676

Submit You have used 1 of 2 attempts

Note: The Gaussian mixture model can be extended to the case where each mixture co
covariance matrix Σ_j. The case that we have studied so far is a special case where $\Sigma_j =$
identity matrix of size $d \times d$. The EM algorithm can also be extended to work in this gen

Useful resource WITH example calculation

https://www.cs.toronto.edu/~rgrosse/csc321/mixture_models.pdf Page 5ff. for posterior calculation... etc.

Should d = 5 in the E step?

Should d = 5 in the E step?