Unit 4 - lecture 16: Mixture Models, EM algorithm

Objectives:

- Review Maximum likelihood Estimation (MLE) of mean and variance in Gaussian statistical model
- Define Mixture Models
- Understand and derive ML estimates of mean and variance of Gaussians in a Observed Gaussian Mixture Model.
- Understand Expectation Maximization (EM) algorithm to estimate mean and variance of Gaussian in an Unobserved Gaussian Mixture Model

MLE for Multinomial and Gaussian Models:

The first distribution, the first generative model that we've seen were multinomials. In this case we assume that we have some set of possible outcomes. If we're talking about language it will be maybe the vocabulary, W. and we would also assume that we have certain likelihood to generate particular whole w in this vocabulary. So it means that the likelihood of generating theta w : θ_w, given parameters θ would be θ_w. $p(w|\theta)=\theta_w$, $\sum_{w \in W} \theta_w = 1$, $\theta_w \geq 0$

$p(D|\theta) = \prod_{w \in W} \theta_w^{n(w)}$ (likelihood of a document D)

The second distribution is called Gaussians, we assume that the particular set of points has a particular center and also will have variance.

$$p(x|\mu,\sigma^2) = N(x;\mu,\sigma^2 I) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|x-\mu\|^2}{2\sigma^2}}$$

The first question that we asked, is how I can actually estimate these parameters? θ? μ? σ^2?

So we estimated parameters using MLE, Maximum likelihood Estimator.

Gaussian Mixture Model: Definition:

When we have many cluster, K: $N(x, \mu^{(j)}, \sigma_j^2)$, $j=1\dots K$ ← mixture components

$p_1 \quad p_K$, $\sum_{j=1}^{K} p_j = 1$ ← mixture weights

$j \sim \text{multinomial}(p_1, \dots p_K)$ I select the cluster

$x \sim p(x|\mu^{(j)}, \sigma_j^2)$

Likelihood of Gaussian Mixture Model :

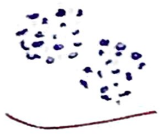
parameters θ : $p_1 \cdots p_K$, $\mu^{(1)} \ldots \mu^{(K)}$, $\sigma_1^2 \ldots \sigma_K^2$

$$p(x|\theta) = \sum_{j=1}^{K} p_j \, N(x, \mu^{j}, \sigma_j^2)$$

$$p(S_n|\theta) = \prod_{i=1}^{m} \sum_{j=1}^{K} p_j \, N(X; \mu^{(j)}, \sigma_j^2)$$

I want to take the derivative in respect of the parameters, make it equal to 0 and find the parameters. It turns out, it's actually a pretty complex task.

So we will start with an easy case called observed :



Someone gave me these points | They gave me the hard assignment. This point belong to the first cluster and so on...

Estimating the Parameters in the Observed Case :

Indicator factor : $\delta(j|i) = \begin{cases} 1, & x^{(i)} \text{ is assigned to } j \\ 0 & \text{otherwise} \end{cases}$ | In an observed case for every point i, there will be just one j to which it belongs.

$$p(S_n|\theta) = \prod_{i=1}^{m} \sum_{j=1}^{K} p_j \, N(X, \mu^{(j)}, \sigma_j^2 I)$$ We need to find these

$$\sum_{i=1}^{m} \left[\sum_{j=1}^{K} \delta(j|i) \log p_j \, N(x^{(i)}, \mu^{(j)} \sigma_j^2 I) \right]$$

for every point we're gonna see to which cluster it belong.

$$\sum_{j=1}^{k} \left[\sum_{i=1}^{m} \delta(j|i) \log p_j \, N(x^{(i)}, \mu^{(j)}, \sigma_j^2 I) \right]$$

The first thing I want to compute is how many members belong to each class, using δ notation to each cluster.

nb of points that belongs to cluster j → $\hat{m}_j = \sum_{i=1}^{m} \delta(j|i)$

Mixture weight for cluster j → $\hat{p}_j = \frac{\hat{m}_j}{m}$

$$\hat{\mu}^{(j)} = \sum_{i=1}^{m} \frac{\delta(j|i) \cdot x^{(i)}}{\hat{m}_j}$$

$$\hat{\sigma}_j^2 = \frac{1}{\hat{m}_j d} \sum_{i=1}^{m} \delta(j|i) \cdot \| x^{(i)} - \mu^{(j)} \|^2$$

The EM Algorithm:

We observe m data points $x_1, \dots x_m$ in $\mathbb{R}^d$. We wish to maximize the GMM likelihood with respect to the parameters set $\theta = \{p_1, \cdots p_K, \mu^{(1)}, \dots, \mu^{(K)}, \dots, \sigma_1^2, \dots \sigma_K^2\}$

Maximizing the log-likelihood $\log\left(\prod_{i=1}^{m} p(x^{(i)}|\theta)\right)$ is not tractable in the setting of GMMs. There is no closed-form solution to finding the parameter set θ that maximize 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 $x^{(i)}$ was generated by cluster j, for every $i = 1, \dots, m$ and $j = 1, \dots, K$

This step assumes the knowledge of the parameter set θ. We find the posterior using the following eq:

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

M-step:

The M step of the algorithm maximizes a proxy function $\hat{\ell}(x^{(1)}, \dots, x^{(m)}|\theta)$ of the log-likelihood over θ, where:

$$\hat{\ell}(x^{(1)}, \dots, x^{(m)}|\theta) \triangleq \sum_{i=1}^{m} \sum_{j=1}^{K} p(j|i) \log\left(\frac{p(x^{(i)} \text{ and } x^{(i)} \text{ generated by cluster } j|\theta)}{p(j|i)}\right)$$

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

$$\ell(x^{(1)}, \dots, x^{(m)}|\theta) = \sum_{i=1}^{m} \log\left[\sum_{j=1}^{K} p(x^{(i)} \text{ generated by cluster } j|\theta)\right]$$

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

$$\hat{\mu}^{(j)} = \frac{\sum_{i=1}^{m} p(j|i)\, x^{(i)}}{\sum_{i=1}^{m} p(j|i)} \qquad \hat{p}_j = \frac{1}{m}\sum_{i=1}^{m} p(j|i) \qquad \hat{\sigma}_j^2 = \frac{\sum_{i=1}^{m} p(j|i)\, \|x^{(i)} - \hat{\mu}^{(j)}\|^2}{d \sum_{i=1}^{m} p(j|i)}$$

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.