AIML - CS 337

Lecture 19: Mixture Models

20/10/2022

Lecturer: Abir De *Scribe: Course Team*

1 Introduction

We observe a data set $D = \{X_i\}_{i=1}^{N}$, where each $X_i = x_i$ is being sampled from one of the K mixture components.

Each of the mixture component is a multivariate Gaussian density with its own parameters $\theta_k = \{\mu_k, \sum_k\}$

$$p_k(x_i|\theta_k) = \frac{1}{(2\pi)^{d/2}|\sum_k|^{1/2}}\boldsymbol{e}^{-\frac{1}{2}(x-\mu_k)^t \sum_k^{-1}(x-\mu_k)}$$

We now have to estimate the parameters of the K mixture components, θ_k and the mixture weights, which represent the probability that a randomly selected $\bar{x}$ was generated by k^{th} component, $\pi_k = P(c(\bar{x}) = k)$, where $\sum_{k=1}^{K} \pi_k = 1$.

2 Computing posterior distribution $P(c(X_i) = k|X_i)$

Using initial estimates for ω, we obtain the posterior in the following way -

$$P(c(X_i) = k \mid X_i, \omega) = \frac{P(\boldsymbol{X_i} \mid c(X_i) = k, \theta_k).P(c = k)}{\sum_m P(\boldsymbol{X_i} \mid c(X_i) = m, \theta_m)P(c = m)} = \frac{N(X_i; \theta_k)\pi_k}{\sum_m N(X_i; \theta_m)\pi_m}$$

This follows a direct application of Bayes rule. These membership weights reflect the uncertainty, given $X_i = x_i$ and ω, about which of the K components generated vector $X_i = x_i$.

3 Maximum Likelihood Estimation

The complete set of parameters for a mixture model with K components is -

$$\omega = \{\pi_1, \pi_2, ..\pi_K, \theta_1, .., \theta_K\}$$

We now maximize the likelihood of data, $P(D) = P(X_1 = x_1, X_2 = x_2, ..., X_N = x_N)$ w.r.t ω.

$$P(D) = \prod_{i=1}^{N} P(\boldsymbol{X_i} = \boldsymbol{x_i})$$

$$\implies \log(P(D)) = \sum_{i=1}^{N} \log(P(\boldsymbol{X_i} = \boldsymbol{x_i}))$$

We know that marginal probability of X_i is,

$$P(\boldsymbol{X}_i = \boldsymbol{x}_i) = \sum_{k=1}^{K} P(\boldsymbol{X_i = x_i} \mid c(X_i) = k)P(c = k)$$

$$\implies P(\boldsymbol{X}_i = \boldsymbol{x}_i) = \sum_{k=1}^{K} P(\boldsymbol{X_i = x_i} \mid c(X_i) = k)\pi_k$$

Using the above,

$$\log(P(D)) = \sum_{i=1}^{N} \log(\sum_{k=1}^{K} P(\boldsymbol{X_i} \mid c(X_i) = k)\pi_k)$$

Differentiating the above w.r.t π_k, μ_k and $\sum_k$, we obtain the new parameters (and using the equation presented in Section 2)-

Let $N_k = \sum_{i=1}^{N} P(c(X_i) = k|X_i, \omega)$

$\pi_k^{new} = \frac{N_k}{N}$

$\mu_k^{new} = (\frac{1}{N_k}) \sum_{i=1}^{N} X_i.P(c(X_i) = k|X_i, \omega)$

$\sum_k^{new} = (\frac{1}{N_k}) \sum_{i=1}^{N} P(c(X_i) = k|X_i, \omega).(X_i - \mu_k^{new}).(X_i - \mu_k^{new})^t$

4 Iterative Procedure for Parameter Estimation

We now work on choosing a suitable initial prior for π_k. Entropy of a distribution is defined as $-\sum_{i=1}^{N} P(\boldsymbol{X_i}) * \log(P(\boldsymbol{X_i}))$ where $\boldsymbol{X_i}$ are random variables of the distribution. In a K-means cluster distribution, we have $\pi_1, \pi_2, .., \pi_K$. In order to maximize the randomness, we assign each one of the random variables probability 1/K.

Now, using the above initial prior for π_k, and some initial parameter estimates θ_k, we derive the posterior $P(c(X_i) = k|X_i)$ (membership weights) as presented in Section 2.

Using these new membership weights, we calculate the new π_k, μ_k and $\sum_k$ using the equations given at the end of Section 3 (derived by differentiating the log likelihood).

Using these new parameter estimates, we calculate the new membership weights and repeat the steps again until the value of likelihood of data converges.

Log likelihood of data - $\log \prod_{i=1}^{N} P(\boldsymbol{X_i}) = \sum_{i=1}^{N} \log(\sum_{k=1}^{K} P(\boldsymbol{X_i} \mid c(\boldsymbol{X_i}) = k) P(c = k))$

Let $P_\omega = P(\boldsymbol{X_i} \mid c(\boldsymbol{X_i}) = k), P_c = P(c = k \mid \boldsymbol{X_i})$

$\omega = \omega^{t-1}$

At time t, $\max_{\omega} \sum_{i=1}^{N} \log(\sum_{k=1}^{K} P_\omega P_c(\omega^{t-1}))$ will give us the new parameter estimates ω

5 Representation in terms of Expectation

We can also represent the likelihood of data $\{\prod_{i=1}^{N} P(\boldsymbol{X_i})\}$ as below.

Now, $P(\boldsymbol{X}) = \sum_{Z} P(X|Z)P(Z)$

implies $P(\boldsymbol{X}) = \mathbf{E}_{\boldsymbol{Z}}[P(\boldsymbol{X}|\boldsymbol{Z})]$

Hence, $P(\boldsymbol{X_i}) = \mathbf{E}_c[P(\boldsymbol{X_i} \mid c)]$

$\prod_{i=1}^{N} P(\boldsymbol{X_i}) = \prod_{i=1}^{N} \mathbf{E}_c[P(\boldsymbol{X_i} \mid c)]$

Now, $\prod_{i=1}^{N} \sum_{k=1}^{K} P(\boldsymbol{X_i}|\boldsymbol{c} = \boldsymbol{k}) P(\boldsymbol{c} = \boldsymbol{k})$

is equal to, $\sum_{k_1=1}^{K} \sum_{k_2=1}^{K} \sum_{k_3=1}^{K} .. \sum_{k_N=1}^{K} (\prod_{i=1}^{N} P(\boldsymbol{X_i}|\boldsymbol{c} = \boldsymbol{k_i}) P(\boldsymbol{c} = \boldsymbol{k_i}))$

$\prod_{i=1}^{N} P(\boldsymbol{X_i}) = \mathbf{E}_{(k_1, k_2, k_3 ..., k_N)}[\prod_{i=1}^{N} P(\boldsymbol{X_i} \mid c = k_i)]$

6 Mixture Model to K-Means iterative algorithm

Entropy of a distribution is defined as $S(X) = \sum_{i=1}^{N} P(X_i) \log(P(X_i))$ where X_i are random variables of the distribution. Entropy is maximised when all of these probabilities are equal (easily proved with differentiation).

So we set the prior $w_k = 1/K$ for all k initially to maximise entropy in K-Means. We set random initial parameter estimates θ. In addition, for later iterations we set $P(c = k) = \boldsymbol{I}(c = k)$ where $\boldsymbol{I}$ is the delta function.

Using maximum likelihood estimation of data, we calculate the new parameters and weights and stop when the likelihood converges.