Lecture 9 : 4th Feb 2026

EM Algorithm :

Recall

ELBO $\quad F_\theta(q) = \mathbb{E}_{q(z)} \log \frac{p_\theta(x,z)}{q(z)}$

the optimal value

$$q(z) = p_\theta(z|x)$$

Note :

$q(z)$ is a fun^{c} of θ

Suppose we evaluate $F_\theta(q)$ at some θ^t

$$F_\theta(q)\Big|_{\theta=\theta^t} = \mathbb{E}_{q(z)} \log \frac{p_\theta(x,z)}{q(z)}$$

with $\quad q(z) = p_{\theta^t}(z|x)$

$q(z)$ and $F_\theta(q)$ can be evaluated at different values of parameter θ

EM :

Initialize $\quad \theta^1$

for $t=1$ to Convergence:

$$q^{t+1}(z) = p_{\theta^t}(z|x)$$

Compute

E-step $\quad F_\theta(q^{t+1}) = \mathbb{E}_{q^{t+1}(z)} \log \frac{p_\theta(x,z)}{q^{t+1}(z)}$

M-step $\quad \theta^{t+1} = \arg\max_\theta F_\theta(q^{t+1})$

End For

Convergence of EM Algorithm

Claim with EM, $l(\theta^{t+1}) \geq l(\theta^t)$

proof: Consider θ^t

$$q^{t+1}(z) = p_{\theta^t}(z|x)$$

$$l(\theta^t) = F_\theta(q^{t+1})\Big|_{\theta=\theta^t}$$

In M step

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}} \; F_\theta(q^{t+1})\Big|_{\theta=\theta^t}$$

$$\therefore \quad F_\theta(q^{t+1})\Big|_{\theta=\theta^{t+1}} \geq F_\theta(q^{t+1})\Big|_{\theta=\theta^t}$$

also $$F_\theta(q^{t+1})\Big|_{\theta=\theta^{t+1}} \leq \ell(\theta^{t+1})$$

By Jensen's inequality

$$\Rightarrow \quad \ell(\theta^{t+1}) \geq \ell(\theta^t) \quad \blacksquare$$

→ likelihood func never decreases when we run EM algo

→ EM can stuck in local minima

→ stopping criteria based on change in the likelihood func

EM for GMMs:

$$p_\theta(x) = \sum_z p_\theta(x, z)$$

$$= \sum_z \alpha_j \, p_{\theta_j}(x)$$

$$p_{\theta_j}(x) \sim \mathcal{N}(x; \mu_j, \Sigma_j)$$

$$= \sum_z p_\theta(z) \, p_\theta(x \mid z)$$

$$z \sim \text{Discrete}\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$$

$$p_\theta(x \mid z = j) \sim \mathcal{N}(x; \mu_j, \Sigma_j)$$

$$\theta = \{\alpha_j, \mu_j, \Sigma_j\}_{j=1}^{m}$$

$\alpha_j \in \mathbb{R}$

$\mu_j \in \mathbb{R}^d$

$\Sigma_j \in \mathbb{R}^{d \times d}$

step 1:

Initialize θ^t

step 2: Compute

$$q^{t+1}(z) = p_{\theta^t}(z|x)$$

$$= \frac{p_{\theta^t}(x|z)\, p_{\theta^t}(z)}{\sum_j p_{\theta^t}(x|z)\, p_{\theta^t}(z)}$$

$$= \frac{N(x; \mu_j^t, \Sigma_j^t)\ \alpha_j^t}{\sum_{j=1}^{m} N(x; \mu_j^t, \Sigma_j^t)\ \alpha_j^t}$$

we can have a closed form soln for this.

step 3: E step: Compute $F_\theta(q^{t+1})$

$$= \mathop{\mathbb{E}}_{q^{t+1}(z)} \log \frac{p_\theta(x,z)}{q^{t+1}(z)}$$

$$= \mathbb{E}_{p_{\theta^t}(z|x)} \log \frac{p_\theta(x,z)}{p_{\theta^t}(z|x)}$$

$$= \sum_{j=1}^{m} \alpha_j^t \log \frac{p_\theta(x|z)\, p_\theta(z)}{p_{\theta^t}(z|x)}$$

$=$

Solve this with proper notations.

Step 4: M step

$$\mu_j^{t+1} = \underset{\mu_j}{\operatorname{argmax}} \; F_\theta(q^{t+1})$$

Solve it .

Lecture 10: 9th Feb 2026:

Bayesian Methods for density Estimation

* Recall in MLE, the parameter set "θ" is considered to be a deterministic variable, that is fully dependent on data.
* When n is small, MLE might not be optimal.

eg: Consider MLE for a Bernoulli RV.

$$\theta_{MLE} = \frac{\#\text{ trial has yeilded "1"}}{n}$$

Suppose $D = \{1, 1, 1, 1\}$

$$\theta_{MLE} = \frac{4}{4} = 1$$

is it a good estimate ?

Can we rellay fully on the data. ?

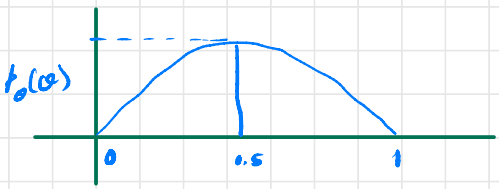
The above estimate is not "optimal", if the true value is around 0.5.

* How do we incorporate the prior belief / information into the estimation technique.

* Assume the parameter "θ" to be a RV & model the prior belief in its distribution.

* Suppose in the previous example of Bernoullie RV.

$p_\theta(\theta)$



* prior belief is completly independent of data

* use this in addition to the Data.

* Suppose $X \sim p_x$

$\theta \sim p_\theta$: prior on parameter.

Compute

$$p_{\theta|x} \propto \underbrace{p_\theta(\theta)}_{\text{prior}} \underbrace{p_\theta(x|\theta)}_{\text{Likelihood}}$$

* How to use $p_{\theta|x}$?

MAP Estimate : Take the Mode of $p_{\theta|x}$ as the estimate for θ.

Conjugate Priors:

* Choice of p_θ is user choice.
* The prior on θ, that ensures the posterior to have the same distributional form as that of the prior.

$p_\theta(\theta)$ is bounded b/w 0 & 1

$p_{\theta|x}$ Can't be Gaussian with infinite Support.

∴ Choose p_θ s.t $p_{\theta|x}$ is also of a same form.

Example of MAP for Bernoullie.

$X \sim$ Bernoullie (θ)

$$p_{x|\theta} = \theta^x (1-\theta)^{1-x}$$

$$p_\theta(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\ \theta^{\alpha-1}\ (1-\theta)^{\beta-1}$$

where $\Gamma()$ is std gamma func & α, β are user choosen params

Compute : (assuming one data point)

$$p_{\theta|x} \propto p_{x|\theta}\, p_{\theta}$$

$$p_{\theta|x} \propto \left[\theta^{x}(1-\theta)^{1-x}\right] \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$\propto \theta^{x+\alpha-1} (1-\theta)^{\beta-x}$$

In MAP estimate, θ_{MAP} is taken to be the mode of $p_{\theta|x}$

* In the above case,

$$\theta_{MAP} = \frac{x+\alpha-1}{\alpha+\beta-2} \Big/ \frac{\sum_{i=1}^{n} x_i + \alpha - 1}{n+\alpha+\beta-2}$$

$$\theta_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

as $n \to \infty$, the belief on the data Vanishes.

$\Rightarrow$ MAP converges to MLE

Non - Parametric Density Estimate

* These methods can evaluate the density func^c at a point without making any parametric func^c form assumption;

* Suppose p_x is the density func^c that is being evaluated.

$$x \in \mathbb{R}^d$$

"R" denote a "region" on the support of p_x

* Let P be the prob of a point "x" falling within the region, R

$$P = \int_{\mathbb{R}} p_x(x)\, dx$$

* Suppose we have "n" datapoints drawn (iid) from p_x

* if the event of a data point falling within the region R, be modeled as a Bernoullie Rv.

* the MLE for that parameter is given by

$$P = \frac{K}{n}$$

where "K" is the number of points with in the region.

* From earlier wkt

$$P = \int_R p_x(x)\, dx$$

$$\therefore \frac{K}{n} = \int_R p_x(x)\, dx.$$

if R is small enough

$$\frac{K}{n} = p_x(x)\, V \quad \text{where}$$

"V" is the Volume of the region.

$$\Rightarrow p_x(x) = \frac{K}{nV}$$

Density is Const over the Volumn

Similar to histogram.

* 2 ways to Construct Non parametric density estimate

  1) Fix V & Count K

  2) Fix K & grow V to encompass K points.

Parzen window Estimates

* Fix V & Comt K to evaluate density
* Suppose we want to evaluate the density fun^{n} at a point x.
* Let R be a d-dimensional hyper cube aramd x with length h

$$V = h^d$$

* to Count "k" define a window fun^{n}

$$\phi(u) = \begin{cases} 1 & \text{if } |u_j| \le \frac{1}{2} \quad \forall j = 1, \dots, d \\ 0 & \text{otherwise} \end{cases}$$

$\phi\left(\frac{x - x_i}{h}\right)$ => is "x_i" is within hyper cube of Volume h^d Centered around x_i:

$$p_x(x) = \frac{k}{nV}$$

$$\hat{p}_x(x) = \frac{\sum_{i=1}^{n} \phi\left(\frac{x - x_i}{h}\right)}{n \cdot h^d}$$

* $\phi()$ is very hard, we need to make it more smoother by choosing

$$\phi(u) = \exp\left(-\| u - u_0 \|_2^2\right)$$

* $\phi()$: kernel, depending on the choice of kernel the estimate changes

Nearest Neighbour estimate :

* non parametric estimate where fix k & grow V.

$$\hat{p}_x(x) = \frac{k}{nV}.$$

Example Use Case : "m" Class Classification.

* Bayes classifier with NN density estimate for $\hat{p}_{y|x}$

* Suppose we want to estimate the class of sample "x"

* we place a Volume V around "x" & capture K- samples

* Let K_i be the number of points within Volume V corresponding to the i^{th} class.

$$\sum_{i=1}^{m} K_i = K$$

* The NN estimate for $p(x, y_i) = \frac{K_i}{nV}$

* now to construct

$$p_{y|x}(y=y_i|x) = \frac{p(x, y_i)}{\sum_{i=1}^{m} p(x, y_i)}$$

$$= \frac{K_i/nV}{\sum_{j=1}^{m} K_j/nV}$$

$$= K_i/K$$

$$h_B(x) = i \quad \text{if} \quad p(y_i|x) > p(y_j|x) \quad \forall j \neq i$$

$$= i \quad \text{if} \quad \frac{k_i}{K} > \frac{k_j}{K} \quad \forall j \neq i$$

$$= i \quad \text{if} \quad k_i > k_j \quad \forall j \neq i$$

This is the K-Nearest neighbour Classifier.

* True risk with NN is atmost twice that of Risk associated with Bayes Classifier. : Proof ??