Lecture 7 : 28th Jan 2026 :

Density Estimation

Recall : $h^*(x) = \underset{h}{\operatorname{argmin}} \left[\underset{\mathbb{P}_{XY}}{\mathbb{E}} \; L(h(x), Y) \right]$

for 0-1 loss,

$$h^*_B(y) = \begin{cases} 1 & \text{if } p_{Y=1|X} > p_{Y=0|X} \\ 0 & \text{otherwise} \end{cases}$$

* Implementing the optimal Bayes classifier demands estimating densities from data

Eg : Given D, Estimate p_X, p_Y, $p_{X|Y}$

Maximum Likelihood (Minimal KL) Estimation :

Given $V = \{ v_1, v_2, \ldots, v_n \} \sim$ iid $\mathbb{P}_V$,

Estimate p_V

Start with p_θ &

$$\theta = \underset{\theta}{\operatorname{argmin}} \; D_{KL}(\mathbb{P}_V \,\|\, \mathbb{P}_\theta)$$

$$\approx \underset{\theta}{\operatorname{argmax}} \left[\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(v_i) \right]$$

$v_i \sim$ iid $\mathbb{P}_v$

Examples of ML Estimation:

a) $p_\theta(v) \sim N(v; \theta, I)$ model

$v \in \mathbb{R}^d$, $\theta \in \mathbb{R}^d$

we have $v_1, v_2 \cdots v_n \sim$ iid $\mathbb{P}_v$

$$p_\theta(v) = \frac{1}{(2\pi)^{d/2} |I|^{1/2}} \exp\left\{ -\frac{1}{2} (v-\theta)^T I (v-\theta) \right\}$$

Consider

$$l(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(v_i)$$

$$\propto \frac{-1}{n} \sum_{i=1}^{n} \| v_i - \theta \|_2^2$$

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \; l(\theta)$$

$$\theta^* = \frac{1}{n} \sum_{i=1}^{n} v_i$$ MLE -

2) Example : Generalized Discrete RV

$D = \{ v_1, v_2, \cdots, v_n \} \sim$ iid P_V

$v_i \in \{ a_1, a_2, \cdots, a_m \}$ with Prob

$\{ p_1, p_2, \cdots, p_m \}$

parameters $\theta = \{ p_1, p_2, \cdots, p_m \}$

To express the mass fun^c for this RV, we need One-hot representation.

One-hot Representation :

For each v_i, define an auxillary RV z_i as follows :

$$z_i = [z_i^1, z_i^2, \cdots, z_i^m], \quad z_i \in \{0,1\}$$

$$z_i^j = \begin{cases} 1 & \text{for } j: \ v_i = a_j \\ 0 & \text{otherwise} \end{cases}$$

Eg : $D = \{ 2, 3, 4, 1, 6 \cdots \}$

$v_i \in \{1,2,3,4,5,6\}$

$z_1 = [0\ 1\ 0\ 0\ 0\ 0]$

$z_2 = [0\ 0\ 1\ 0\ 0\ 0]$

New data $D = \{z_1, z_2, \cdots, z_n\}$

$z_i \in \{0,1\}^m$

$$p_\theta(z_i) = \prod_{j=1}^{m} p_j^{z_i^j} \qquad \theta = \{p_1, p_2, \cdots, p_m\}$$

$$\theta^* = \underset{\theta}{\text{argmax}} \ \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(z_i)$$

$$= \underset{\theta}{\text{argmax}} \ \frac{1}{n} \sum_{i=1}^{n} \log \prod_{j=1}^{m} p_j^{z_i^j}$$

$$= \underset{\theta}{\text{argmax}} \ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} z_i^j \log p_j$$

adding constraints to the above problem

$$0 \le p_j \le 1 \quad \& \quad \sum_{j=1}^{m} p_j = 1$$

New objective

$$\underset{\theta}{\text{argmax}} \ l(\theta)$$

$$\text{s.t} \quad 0 \le p_j \le 1 \quad \& \quad \sum_{j=1}^{m} p_j = 1$$

writing the lagrangian

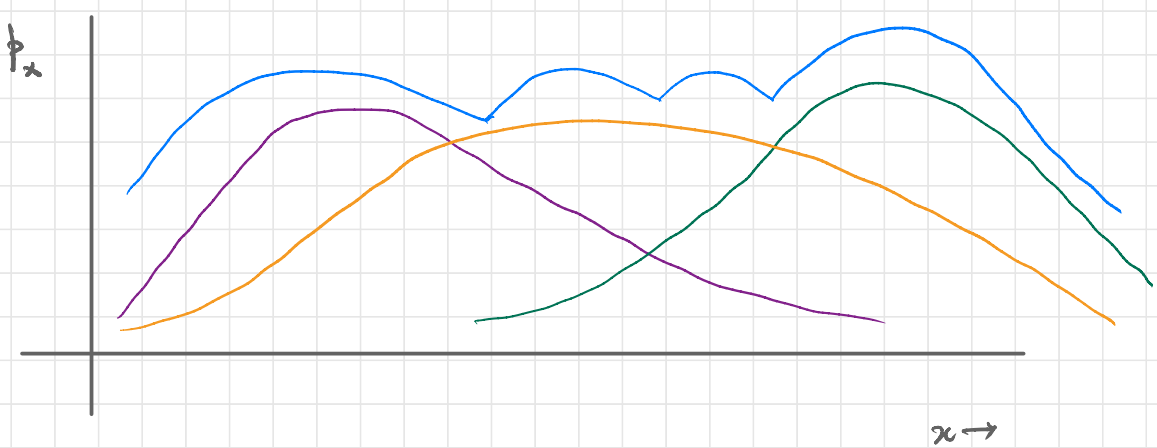
$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} z_i^j \log p_j + \lambda\left(\sum_{j=1}^{m} p_j - 1\right)$$
$$+ \alpha\left(p_j - 1\right) + \beta\left(-p_j\right)$$

after solving the above problem

$$p_j^* = \frac{\#\ a_j \text{ as occured}}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} z_i^j$$

Density Estimation for Mixture Distributions:



Multi-model densities are not "well-estimated" with uni-model models.

Soln: Try a multi-modal model.

Mixture density fun^{c}:

Suppose $v \in \mathbb{R}^d$

Define a mixture density

$$p_\theta(v) = \sum_{j=1}^{m} \alpha_j \, p_{\theta_j}(v)$$

$\alpha_j \in [0,1]$ $\quad p_{\theta_j}$: density fun^{c}

Gaussian Mixture models: GMM.

$$p_{\theta_j} \sim N(\cdot\,; \theta_j)$$

GMMs are Universal density approximations for CRVs.

Lecture 8 : 2nd Feb 2026

Mixtures Densities & EM

Recall

A mixture density model

$$p_\theta(v) = \sum_{j=1}^{m} \alpha_j \, p_{\theta_j}(v)$$

where $p_{\theta_j}(v)$ is a density func

$$0 \le \alpha_j \le 1 \quad \& \quad \sum_{j=1}^{m} \alpha_j = 1$$

Gaussian Mixture Model

$$p_{\theta_j}(\cdot) \sim N(\cdot ; \mu_j, \Sigma_j) \qquad j = 1, \dots, m$$

MLE for mixture density

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log p_\theta(v_i)$$

in the case of mixture densities.

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

$$\sum_{i=1}^{n} \log p_\theta(x_i) = \sum_{i=1}^{n} \log\left(\sum_{j=1}^{m} \alpha_j \, p_{\theta_j}(x_i)\right)$$

log of Sums

exp will not get cancelled with log

∴ we need an iterative algorithm.

Latent Variable Models:

Given $D = \{x_1, x_2, \cdots, x_n\} \sim$ i.i.d $\mathbb{P}_x$

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

Introduce into the system another RV into the system which is un-observed.

→ this is known as hidden / latent RV. : z

Mathematically

$\forall\ x_i \in \mathbb{R}^d$, define $z_i \in \mathbb{R}^k$, typically $k << d$

→ x & z are assumed to be correlated.

→ z can be CRV @ DRV.

With this latent variable model is defined as

$$p_\theta(x) = \sum_z p_\theta(x, z) \quad \text{if } Z \text{ is DRV}$$

$$p_\theta(x) = \int_z p_\theta(x, z)\, dz \quad Z \text{ is CRV.}$$

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

} All these exists

MLE for Latent Variable Models:

→ For a latent variable models, both the model parameters & the density over the latent variable needs to be estimated.

* we have the Likelihood fun^c $l(\theta)$

Note: we have considered one data point

$$l(\theta) = \log p_\theta(x)$$

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

Assuming Z is DRV.

Suppose $q(z)$ is some density over z.

$$\ell(\theta) = \log \sum_z \left[p_\theta(x,z) \frac{q(z)}{q(z)} \right]$$

$$= \log \sum_z q(z) \underbrace{\left(\frac{p_\theta(x,z)}{q(z)} \right)}_{\text{Func of } z}$$

$$= \log \mathop{\mathbb{E}}_{q(z)} \left[\frac{p_\theta(x,z)}{q(z)} \right]$$

Using Jensen's inequality & noting that

log is a concave func^n

$f(\mathbb{E}_x) \leq \mathbb{E} f(x)$: $f(x)$ is convex.

$\therefore$ we have

$$\log \mathop{\mathbb{E}}_{q(z)} \left(\frac{p_\theta(x,z)}{q(z)} \right) \geq \mathop{\mathbb{E}}_{q(z)} \log \left[\frac{p_\theta(x,z)}{q(z)} \right]$$

$$\ell(\theta) \geq F_\theta(q)$$

* $F_\theta(q)$ is now computable $\because$ it has "Sum of log" type terms.

* $F_\theta(q)$ is a lower bound on $l(\theta)$

$l(\theta)$: log Likelihood @ Evidence.

∴ Evidence lower bound (ELBo)

q : Variational distribution

New optimization problem :

$$\theta^* = \underset{\theta}{\text{argmax}}\ l(\theta)$$

$$\hat{\theta}^*, q = \underset{\theta, q}{\text{argmax}}\ F_\theta(q) \quad \because \text{modified optimization}$$

Question : What choice of $q(z)$ would make ELBo tight ?

Consider

$$l_\theta - F_\theta(q) = \log p_\theta(x) - \sum_z q(z) \log \frac{p_\theta(x,z)}{q(z)}$$

$$= \log p_\theta(x) - \sum_z q(z) \log \frac{p_\theta(x)\, p_\theta(z|x)}{q(z)}$$

$$= \log p_\theta(x) - \log p_\theta(x) - \sum_z q(z) \log \frac{p_\theta(z|x)}{q(z)}$$

$$= \sum_z q(z) \log \frac{q(z)}{p_\theta(z|x)}$$

$$= D_{KL}(q(z) \,||\, p_\theta(z|x))$$

Now $F_\theta(q) = \ell(\theta)$ by construction.

$\ell(\theta) - F_\theta(q) = 0$ iff $D_{KL}(q(z) \,||\, p_\theta(z|x)) = 0$

$\therefore$ $q^*(z) = p_\theta(z|x)$

makes the ELBO tight (exactly same)

* if we know $p_\theta(z|x)$ then we need to optimize the modified optimization only on θ

An iterative Algorithm for optimizing $\ell(\theta)$:

1) Initialize θ^0

2) For $t = 1$ to Convergence

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

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

Expectation Step

Estimate

$$\theta^{t+1} \leftarrow \operatorname{argmax} \left[F_{\theta^t}(q^{t+1}) \right]$$

Maximization Step

To show EM guarantees

$$l(\theta^{t+1}) \geq l(\theta^t)$$