5.5 Maximum Likelihood Estimation.

Motivation:

We want some principle from which we can derive specific functions that are good estimators for different models.

The most common: $\boxed{\text{maximum likelihood principle}}$

$X = \{x^{(1)}, \dots, x^{(m)}\}$, m examples drawn iid w.r.t. some unknown true data generating dist. $p_{data}(x)$.

$\{p_{model}(x;\theta)\}_\theta$: a family of prob. dist. indexed by θ.

Def

The $\boxed{\text{maximum likelihood estimator}}$ for θ is defined as

$$\boxed{\theta_{ML}} = \arg\max_\theta \, p_{model}(X;\theta)$$

$$= \arg\max_\theta \prod_{i=1}^m p_{model}(x^{(i)};\theta) \quad \text{(due to iid)}$$

$$= \arg\max_\theta \sum_{i=1}^m \log(p_{model}(x^{(i)};\theta))$$

$$= \arg\max_\theta \frac{1}{m}\sum_{i=1}^m \log(p_{model}(x^{(i)};\theta))$$

$$= \arg\max_\theta \, E_{x\sim\hat{p}_{data}} \log(p_{model}(x;\theta))$$

(where $\hat{p}_{data}$ is the empirical distribution)

$$= \arg\min_\theta \underbrace{-E_{x\sim\hat{p}_{data}} \log(p_{model}(x;\theta))}_{\boxed{\text{cross entropy}}\text{ b/w } \hat{p}_{data}\ \&\ p_{model}}$$

$$= \arg\min_\theta \underbrace{E_{x\sim\hat{p}_{data}} [\log \hat{p}_{data}(x;\theta) - \log p_{model}(x;\theta)]}_{\boxed{D_{KL}(\hat{p}_{data} \| p_{model})}\text{, the } \boxed{\text{KL divergence}}}$$

Rmk:

① In view of KL divergence (as measuring difference), maximum likelihood $\Leftrightarrow$ trying to make p_{model} match $\hat{p}_{data}$.

② maximum likelihood

$\Leftrightarrow$ minimization of $\boxed{\text{negative log-likelihood}}$ ($\boxed{\text{NLL}}$)

$\Leftrightarrow$ - - - - - - - - - - cross entropy. / KL divergence.

5.5.1 Conditional Log-Likelihood and Mean Square Error.

Def

X: all our inputs, Y: all our observed targets.

The $\boxed{\text{conditional maximum likelihood estimator}}$ is

$$\theta_{ML} = \arg\max_\theta P(Y|X;\theta).$$

Rmk:

For iid drawn X, Y,

$$\theta_{ML} = \arg\max_\theta \sum_{i=1}^m \log P(y^{(i)}|x^{(i)};\theta)$$

Prop

If the model

$$p(y|x;\theta) = \underbrace{\mathcal{N}}_{\text{Gaussian}}(y; \underbrace{\hat{y}(x;w)}_{\text{mean}}, \underbrace{\sigma^2}_{\text{variance}}), \text{ w/ } \sigma^2 \text{ fixed}$$

and w as parameter, (so $\theta = w$).

then the maximum likelihood estimator is the same as minimization of MSE on $\{\hat{y}(x;w)\}_w$.

<Pf>

$$\theta_{ML} (= w_{ML})$$

$$= \arg\max_w \sum_{i=1}^m \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y^{(i)} - \overbrace{\hat{y}(x^{(i)};w)}^{\text{denoted } \hat{y}^{(i)}})^2}{2\sigma^2}}\right)$$

$$= \arg\max_w \left(-m\log\sigma - \frac{m}{2}\log(2\pi) - \sum_{i=1}^m \frac{|y^{(i)} - \hat{y}^{(i)}|^2}{2\sigma^2}\right)$$

$$= \arg\min_w \left(\sum_{i=1}^m \frac{|y^{(i)} - \hat{y}^{(i)}|^2}{m}\right) = \arg\min_w \text{MSE}. \quad \#$$

Rmk:

In short, para. modelling the mean of Gaussians w/ fixed variance together maximum likelihood

$\Leftrightarrow$ para. modelling w/ MSE as cost function.