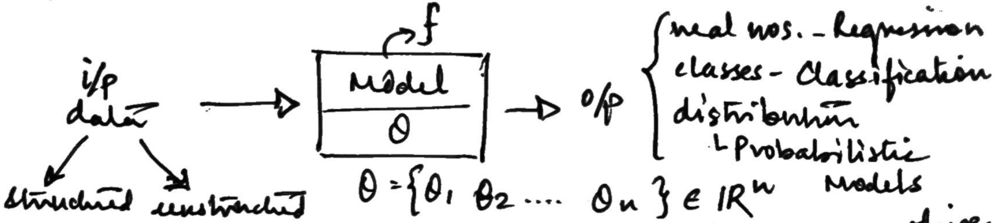
04/08/2025 - AIL7024

→ function

f_θ

↳ parameters of the function

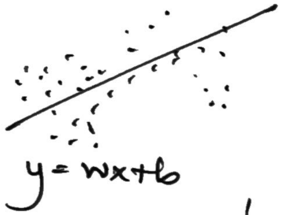


$\theta = \{\theta_1\ \theta_2 \dots \theta_n\} \in \mathbb{R}^n$

$\boxed{f} = wx+b \Rightarrow \theta = \{w, b\}$

↳ convex function

↳ good fit.



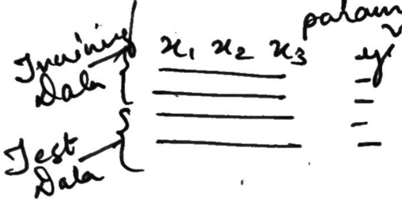
→ one setting

$\theta \in \Theta$

↳ set of all possible parameter values

Maximum Likelihood Estimate (MLE)

Data: $\{(x_i, y_i)\}_{i=1}^{N}$



Model: $f_\theta(x) = y$, linear → $f_\theta(x) = w^T x + b$

↳ $\theta = \{w, b\}$

$x[\ ]\quad w[\ ]$

y: ground truth

$\hat{y}$: prediction/generated output

Loss function $l(y, \hat{y}) = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ ← convex

$y - \hat{y}(x)$ - individual terms can be (-)ve

$$\boxed{\hat{\theta} \text{ or } \theta^* = \underset{\theta}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} l(y_i, \hat{y}_i)}$$

Linear Gaussian Model: $\mathcal{N}(w^T x + b, \sigma^2)$ ← distribution

$P_\theta(y|x) = \mathcal{N}(y; w^T x + b, \sigma^2)$ ← probability value.

MLE - Non ML perspective

$\mathcal{D} = \{x_1, x_2, \dots, x_n\}$ → assume a set of distribution x

$P_\theta : \theta \in \Theta$

assume that $\mathcal{D}$ was sampled from a member of this family

Goal: recover $\hat{\theta}$

likelihood of data

MLE: $\theta_{MLE} = \underset{\theta}{\arg\max}\, P_\theta(\mathcal{D}) = \underset{\theta}{\arg\max} \prod_{i=1}^{N} P_\theta(x_i)$

$$\{\mu_{MLE}, \sigma^2_{MLE}\} = \underset{\mu,\sigma^2}{\operatorname{argmax}} \prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma^2)$$

Taking log, (monotonic fn ; maximizing log = maximizing L(.)),

$$\underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log \mathcal{N}(x_i; \mu, \sigma^2)$$

or, $$\underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log\left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right)\right]$$

or, $$\underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}$$

or, $$\underset{\mu,\sigma^2}{\operatorname{argmax}} -\frac{N}{2\sigma^2} \sum_{i=1}^{N} (x_i-\mu)^2 - \frac{N}{2}\log\sigma^2 = F$$

Now,

$$\frac{\partial F}{\partial \mu} = \sum_i \frac{(x_i-\mu)}{\sigma^2} = 0 \Rightarrow \frac{\sum x_i}{\sigma^2} - \frac{N\mu}{\sigma^2} = 0$$

or, $$\frac{\sum_i x_i}{\cancel{\sigma^2}} = \frac{N\mu}{\cancel{\sigma^2}}$$

or, $$\boxed{\mu = \frac{\sum x_i}{N}.}$$

$$\frac{\partial F}{\partial \sigma^2} = \sum_i \frac{(x_i-\mu)^2}{2\sigma^2} - \frac{N}{2}\frac{1}{\sigma^2} = 0$$

or, $$\boxed{\sigma^2 = \frac{\sum_i (x_i-\mu)^2}{N}}$$

(Shannon's)

Entropy : no. of bits needed to encode information.

$$H(p) = -\sum p_i \overbrace{\log p_i}$$

$$\mathbb{E}[x] \overset{N\to\infty}{\approx} \frac{1}{N}\sum_i x_i$$ infinite data

Cross entropy (CE)

$$H(p,q) = -\sum p_i \log q_i = \mathbb{E}_{p}[-\log q]$$

ground truth distribution

generated distribution

$$= \frac{1}{N}\sum_{i=1}^{N} -\log(q(x_i))$$

using Monte Carlo estimate

↑ MLE = ↓ CE

D_{KL} (KL Divergence) $\Rightarrow D_{KL}(p\|q) = H(p,q) - H(p)$