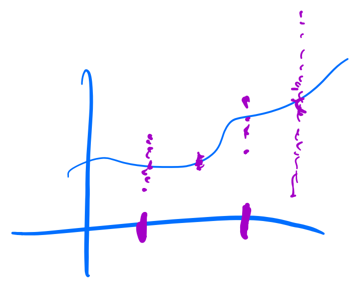
Maximum Likelihood Estimation (MLE)

How about using a different learner from ERM?

$$f_{Bayes} = \underset{f \in \{f \mid f: \mathcal{X} \to \mathcal{Y}\}}{\operatorname{argmin}} L(f) \qquad \underset{\mathbb{P}_{\vec{X},Y}}{E}\left[\ell(f(\vec{x}), Y)\right]$$



$$f_{Bayes}(\vec{x}) = E[Y \mid \vec{X} = \vec{x}]$$

$$= \int_y y \cdot p_{Y|\vec{x}}(y \mid \vec{x})\, dy$$

Let's use the dataset D to estimate $P_{Y|\vec{x}}$

MLE Basics

$D = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$, $\quad \mathbb{P}_D$, p_D

Z_i are i.i.d. with $\mathbb{P}_Z$ and pmf or pdf p_Z

If we have a fixed dataset $D = (z_1, \ldots, z_n)$

how can learn what p is

$\Rightarrow$ pick the p that makes the data most likely

$$p_D(D) = p_D(z_1, \dots, z_n) \overset{\text{independent}}{=} p_{z_1}(z_1)\, p_{z_2}(z_2) \dots p_n(z_n)$$

$$\overset{\text{identically distributed}}{=} p_z(z_1) \cdots p(z_n) \longrightarrow \text{i.i.d.}$$

$$p_{MLE} = \underset{p \in \mathcal{H}}{\arg\max} \prod_{i=1}^{n} p(z_i) \approx p_z \qquad p_{\vec{x},y}$$

We pick the pmf/pdf that makes the data under this prob. dist. most likely.

Ex: $Z_i \sim \text{Bernoulli}(\alpha^*)$ is the i-th flip of an unfair coin

$$p_Z(z) = (\alpha^*)^z (1-\alpha^*)^{(1-z)}$$

$$\mathcal{H} = \{ p \mid p : \mathcal{Z} \to [0,1] \text{ and } p(z) = \alpha^z (1-\alpha)^{(1-z)}, \ \alpha \in [0,1] \}$$

Any $p_\alpha \in \mathcal{H}$ has the form

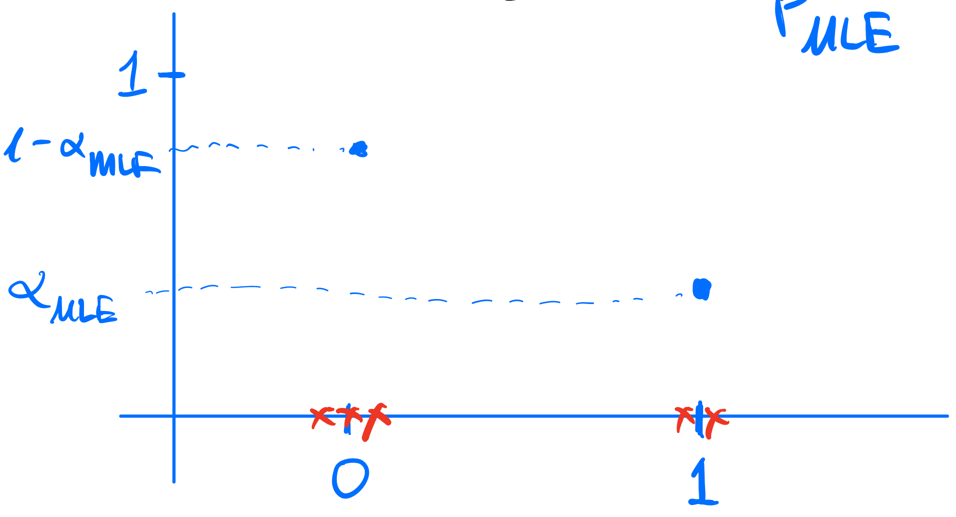
$$p_\alpha(z) = \alpha^z (1-\alpha)^{(1-z)} \text{ for some } \alpha \in [0,1]$$

$$p_{MLE} = p_{\alpha_{MLE}} \approx p_z$$

$$\text{where } \alpha_{MLE} = \underset{\alpha \in [0,1]}{\arg\max} \overbrace{\prod_{i=1}^{n} p_\alpha(z_i)}^{\text{likelihood} = p_D(D \mid \alpha)}$$

$$= p(z_i \mid \alpha)$$

$n=5$ $\quad P_{MLE}$



$\underline{\underline{Ex}}$: $Z_i \sim \mathcal{N}(\mu^*, 1)$ is the i-th persons height

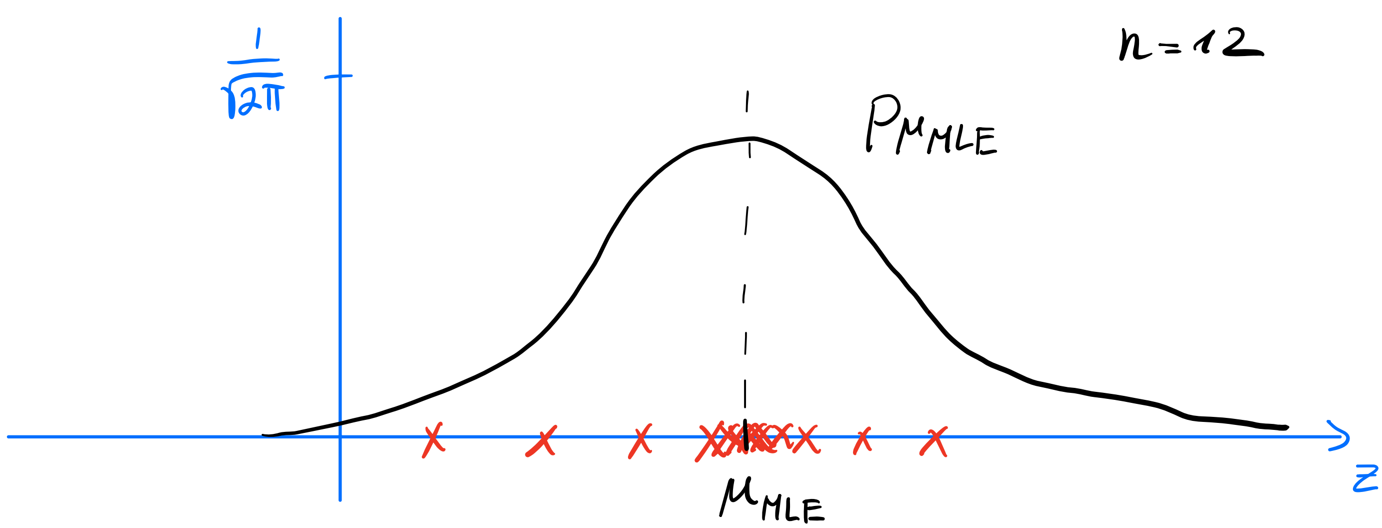
$$p_z(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu^*)^2}{2}\right)$$

$$\mathcal{H} = \left\{ p \mid p: \mathcal{Z} \to [0,\infty) \text{ and } p(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\mu)^2}{2}\right), \mu \in \mathbb{R} \right\}$$

$$P_{MLE} = P_{\mu_{MLE}} \approx P_z$$

where $\mu_{MLE} = \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \prod_{i=1}^{n} \underbrace{P_\mu(z_i)}_{= p(z_i \mid \mu)}$ $\qquad$ and $p(\cdot \mid \mu) \in \mathcal{H}$

all possible inputs



Calculating μ_{MLE}:

$$\mu_{MLE} = \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \prod_{i=1}^{n} p(z_i \mid \mu)$$

have the same maximizer

→ log is monotonically increasing

$\log = \log_e = \ln$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \log\left(\prod_{i=1}^{n} p(z_i \mid \mu)\right)$$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmax}} \sum_{i=1}^{n} \log\left(p(z_i \mid \mu)\right)$$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} \underbrace{- \sum_{i=1}^{n} \log\left(p(z_i \mid \mu)\right)}_{\text{negative log-likelihood} \, = \, -\log p(D \mid \mu)}$$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} - \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z_i - \mu)^2}{2}\right)\right)$$

$\log(a \cdot b) = \log(a) + \log(b)$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} - \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) + \log\left(\exp\left(-\frac{(z_i - \mu)^2}{2}\right)\right)\right]$$

$\log(\exp(x)) = x$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} - \sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) - \frac{(z_i - \mu)^2}{2}\right]$$

$$= \underset{\mu \in \mathbb{R}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \frac{(z_i - \mu)^2}{2}}_{g(\mu)}$$

$$\frac{d}{d\mu} g(\mu) = \frac{-2}{2} \sum_{i=1}^{n} (z_i - \mu) = -\sum_{i=1}^{n} (z_i - \mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \mu = \sum_{i=1}^{n} z_i$$

$$n\mu = \sum_{i=1}^{n} z_i$$

$$\Rightarrow \mu_{MLE} = \mu = \frac{1}{n} \sum_{i=1}^{n} z_i$$ $\rightarrow$ * sample mean (but with $\frac{1}{n}$ instead of $\frac{1}{n-1}$)

If you want to find the pdf of a normal dist that makes the data most likely, calculate the sample mean* over the data.

Estimating $p_{Y|\vec{X}}$

$$D = \left((\vec{X}_1, Y_1), \dots, (\vec{X}_n, Y_n) \right) \in (\mathcal{X} \times \mathcal{Y})^n, \ \mathbb{P}_D, \ p_D$$

$(\vec{X}_i, Y_i)$ are i.i.d with $\mathbb{P}_{\vec{X},Y}$ and $p_{\vec{X},Y}$

$$p_{\vec{x},y}(\vec{x},y) \overset{\text{product rule}}{=} p_{y|\vec{x}}(y \mid \vec{x})\, p_{\vec{x}}(\vec{x})$$

Assume $Y_i \mid \vec{X}_i = \vec{x}_i \sim \mathcal{N}(\vec{x}_i^{\,T} \vec{w}^{*}, 1)$