Important Announcements and Notes (Oct 31)

- Use latest version of lecture notes
- Index k instead of j for derivative
- Estimating p_z (not approximating)

Maximum Likelihood Estimation (MLE)

How about using a different learner from ERM?

$$f_{Bayes} = \underset{f \in \{f \mid f: \mathcal{X} \to \mathcal{Y}\}}{\arg\min} \; L(f) \longrightarrow \mathbb{E}[\ell(f(\vec{X}), Y)]$$

based on $\mathbb{P}_{\vec{X},Y}$

assume squared loss and Regression

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

$$= \int_{\mathcal{Y}} y \, p_{Y|\vec{X}}(y \mid \vec{x}) \, dy$$

Lets use the dataset D to estimate $p_{Y|\vec{X}}$

MLE Basics

$D = (Z_1, \ldots, Z_n) \in \mathcal{Z}^n$, $\mathbb{P}_D$, 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 we estimate p_Z?

Pick p_{MLE} that makes the data the most likely

$$p_D(D) = p_D(z_1, \ldots, z_n) \overset{\text{independent}}{=} p_{z_1}(z_1) \cdots p_{z_n}(z_n)$$

$$\overset{\text{identically distributed}}{=} p_z(z_1) \cdots p_z(z_n)$$

$$= \prod_{i=1}^{n} p_z(z_i)$$

we want: $p_{MLE} = \underset{p \in \mathcal{H}}{\operatorname{argmax}} \prod_{i=1}^{n} p(z_i)$

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

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

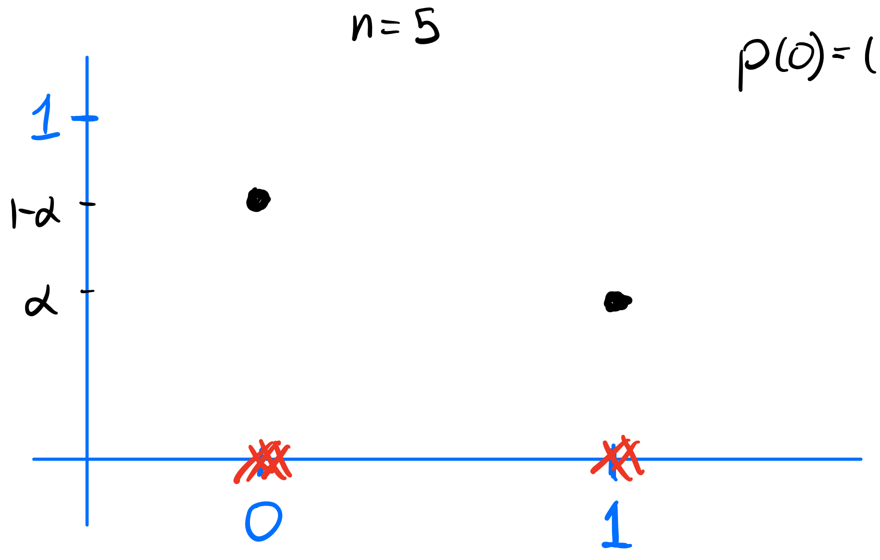
$$\mathcal{H} = \{ p \mid p : \mathcal{Z} \to [0,1] \text{ and } p(z) = \alpha^z (1-\alpha)^{(1-z)}, \ \alpha \in [0,1] \}$$

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

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

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

$$\alpha_{MLE} = \underset{\alpha \in [0,1]}{\operatorname{argmax}} \overbrace{\prod_{i=1}^{n} \underbrace{p_\alpha(z_i)}_{p(z_i \mid \alpha)}}^{\text{likelihood} \, = \, p(D \mid \alpha)}$$



$p(0) = (1-\alpha)$, $p(1) = \alpha$

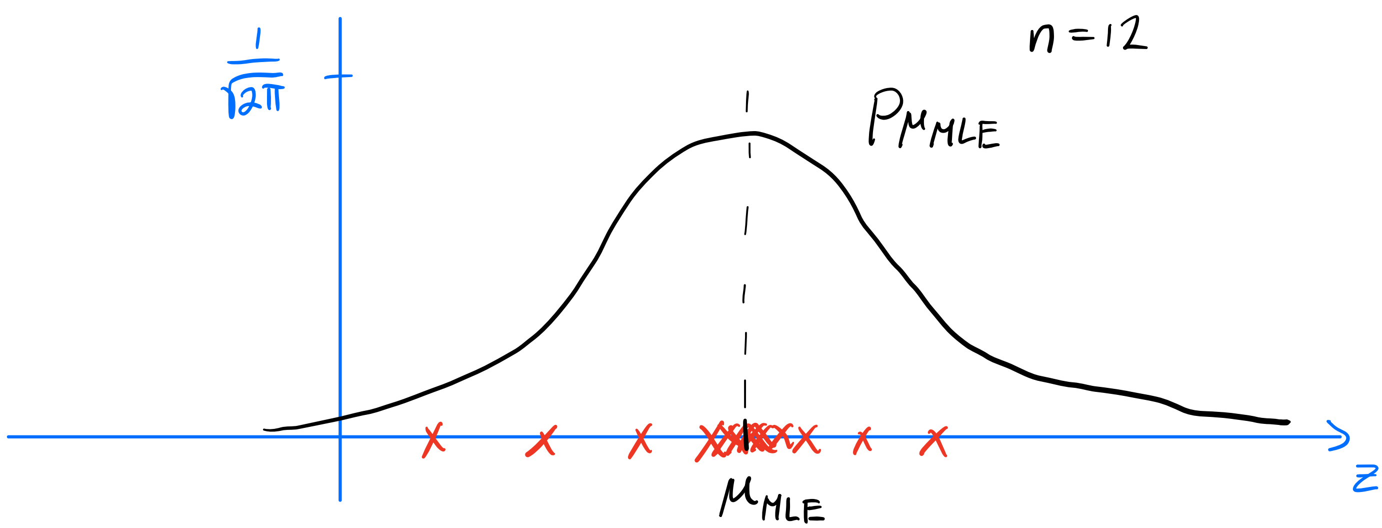
Ex: $Z_i \sim \mathcal{N}(\mu^*, 1)$ is the i-th persons height

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

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

$$p_{MLE} = p_{\mu_{MLE}} \approx p_Z$$

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



Calculating μ_{MLE}:

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

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

$\log = \log_e = \ln$

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

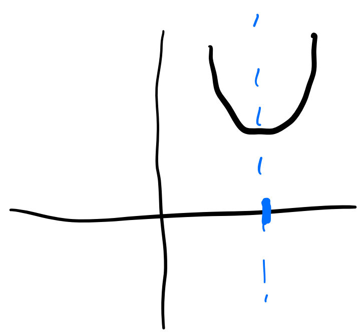
$\log(ab) = \log(a) + \log(b)$

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

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

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

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



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

$$-\sum_{i=1}^{n} z_i + \sum_{i=1}^{n}\mu = 0 \Rightarrow n\mu = \sum_{i=1}^{n} z_i$$

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

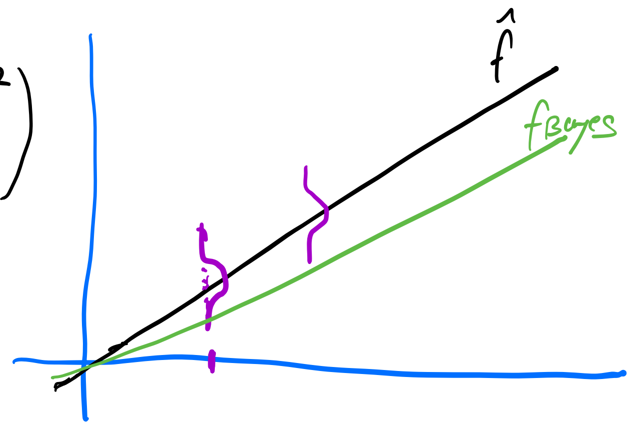
Estimating $P_{Y|\vec{X}}$

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

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

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

$$p_{Y|\vec{X}=\vec{x}}(y|\vec{x}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \vec{x}^T\vec{w}^*)^2}{2}\right)$$



product rule

$$p_{\vec{x},Y}(\vec{x}, y) \overset{\downarrow}{=} p(y|\vec{x})\, p(\vec{x})$$

Calculating $\vec{w}_{MLE}$:

$$\vec{w}_{MLE} = \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\max} \prod_{i=1}^{n} p(\vec{x}_i, y_i | \vec{w})$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} -\log\left(\prod_{i=1}^{n} p(\vec{x}_i, y_i | \vec{w})\right)$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} \sum_{i=1}^{n} -\log\left(p(\vec{x}_i, y_i | \vec{w})\right)$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} \sum_{i=1}^{n} -\log\left(p(y_i | \vec{x}_i, \vec{w})\, p(\vec{x}_i)\right)$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} -\sum_{i=1}^{n} \left[\log\left(p(y_i | \vec{x}_i, \vec{w})\right) + \log\left(p(\vec{x}_i)\right)\right]$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} -\sum_{i=1}^{n} \log\left(p(y_i | \vec{x}_i, \vec{w})\right)$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} -\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - \vec{x}^T \vec{w})^2}{2}\right)\right)$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} -\sum_{i=1}^{n} \left[\log\left(\frac{1}{\sqrt{2\pi}}\right) + \log\left(\exp\left(-\frac{(y - \vec{x}^T \vec{w})^2}{2}\right)\right)\right]$$

$$= \underset{\vec{w} \in \mathbb{R}^{d+1}}{\arg\min} \underbrace{\sum_{i=1}^{n} \frac{\overbrace{(y_i - \vec{x}_i^T \vec{w})^2}^{(\vec{x}_i^T \vec{w} - y_i)^2}}{2}}_{= \frac{n}{2} \hat{L}(\vec{w})}$$

$$\vec{w}_{MLE} = A^{-1} \vec{b} = \hat{\vec{w}}$$

$$P_{MLE}(y | \vec{x}) = p(y | \vec{x}, \vec{w}_{MLE}) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - \vec{x}_i^T \vec{w}_{MLE})^2}{2}\right)$$

$$\approx P_{Y | \vec{X} = \vec{x}}(y | \vec{x})$$

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

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

$$\approx \int_y y \, p(y \mid \vec{x}, \vec{w}_{MLE}) \, dy$$

$$= \mathbb{E}[Y' \mid \vec{X} = \vec{x}] \quad \text{where } Y' \mid \vec{X} = \vec{x} \sim \mathcal{N}(\vec{x}^T \vec{w}_{MLE}, 1)$$

$$= \vec{X}^T \vec{w}_{MLE}$$

$$= \vec{X}^T \hat{\vec{w}}$$

$$= \hat{f}$$