Parametric Methods

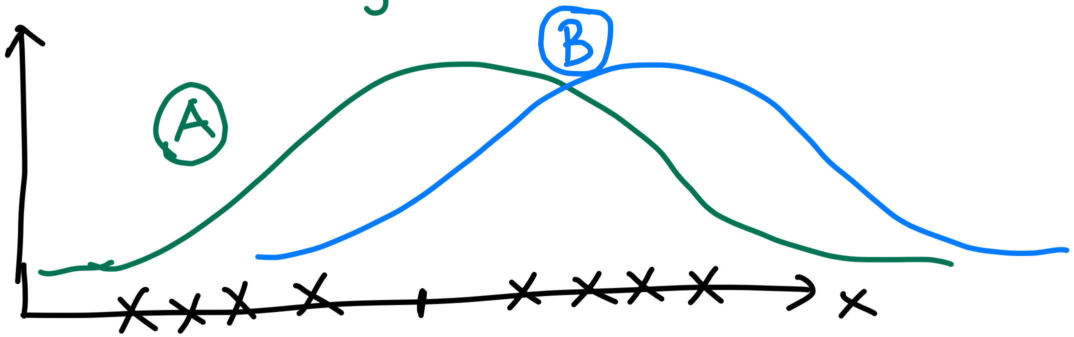
Density Estimation

$$\mathcal{X} = \{x_i\}_{i=1}^{N}$$

N data points
N samples

$x_i \sim p(x_i) \quad \forall i \Rightarrow$ probability distribution $\Downarrow$ parameters(?)

learning these parameters over training data

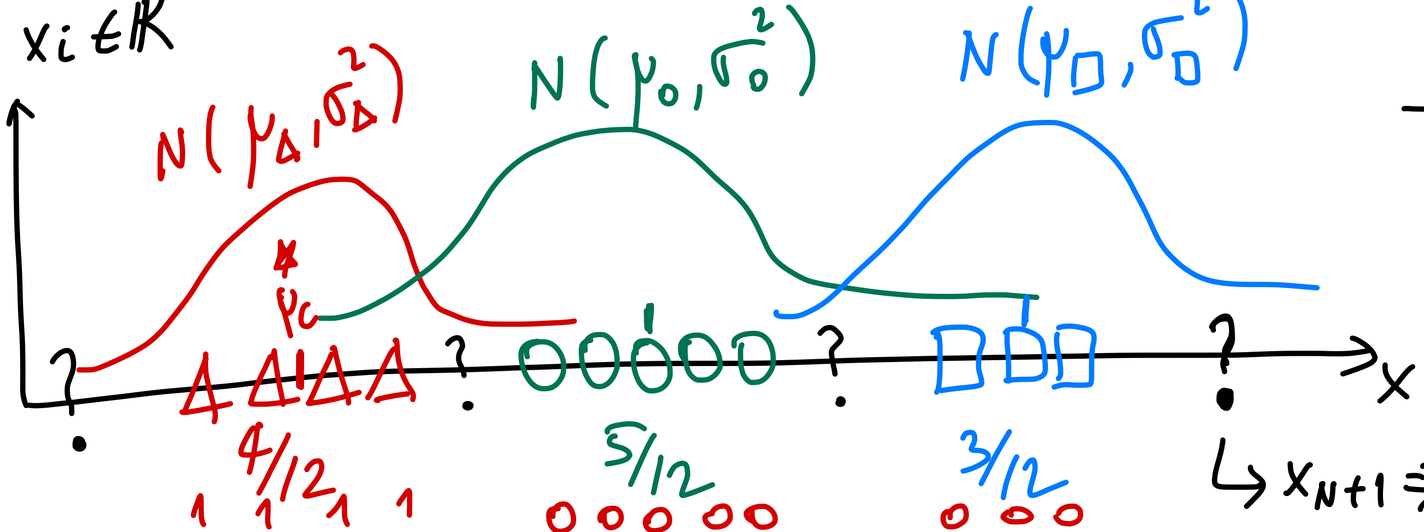


$x_i \sim N(x_i; \mu, \sigma^2)$

$\mu \Rightarrow \hat{\mu}$

$\sigma^2 \Rightarrow \hat{\sigma}^2$

$x_i \in \mathbb{R}$



← boundary moves to the left.

$\hookrightarrow x_{N+1} \Rightarrow y_{N+1} = ?$

$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$ $\quad x_i \in \mathbb{R}$ $\quad y_i \in \{1, 2, 3\}$

$\triangle \quad \bigcirc \quad \square$

class conditional densities $\Rightarrow p(x \mid y=c)$

prior probabilities $\Rightarrow \Pr(y=c)$

BAYES RULE $\quad P(B \mid A) = \dfrac{P(A, B)}{P(A)} = \dfrac{P(A \mid B)\, P(B)}{P(A)}$

$B \Rightarrow y=c$

$A \Rightarrow x$

$$\underbrace{\Pr(y=c \mid x)}_{\text{posterior}} = \frac{p(x \mid y=c)\, \Pr(y=c)}{p(x)}$$

ignored

$x_{N+1} \longrightarrow \Pr(y=c \mid x_{N+1})$

$\Pr(y=1 \mid x_{N+1})$

$\Pr(y=2 \mid x_{N+1})$

$\Pr(y=3 \mid x_{N+1})$

Pick the maximum one

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

$\mathcal{X} = \{x_i\}_{i=1}^{N}$ $\quad x_i \sim p(x_i | \theta) \quad \forall i$

↳ unknown parameters of p

x_i's are assumed to be i.i.d

i → independently

i → identically

d → distributed

$$\text{Likelihood} \equiv p(x_1, x_2, \ldots, x_N | \theta) \Rightarrow \text{full joint}$$

$$\mathcal{L}(\theta | \mathcal{X}) \equiv p(x_1 | \theta)\, p(x_2 | \theta) \cdots p(x_N | \theta)$$

$$\equiv \prod_{i=1}^{N} \boxed{p(x_i | \theta)}$$

$$\log \mathcal{L}(\theta | \mathcal{X}) = \log\left[\prod_{i=1}^{N} p(x_i | \theta)\right]$$

$$= \sum_{i=1}^{N} \log\left[p(x_i | \theta)\right]$$

$$\theta^* = \arg\max_{\theta} \mathcal{L}(\theta | \mathcal{X})$$

$$\theta^* = \arg\max_{\theta} \log \mathcal{L}(\theta | \mathcal{X})$$

$\boxed{\log(a^b) = b.\log(a)}$

$\boxed{\log(a.b) = \log(a) + \log(b)}$

Bernoulli Density

$0 < \pi < 1$ success probability

$$\frac{\partial \log(x)}{\partial x} = \frac{1}{x} \qquad \frac{\partial \log(1-x)}{\partial x} = -\frac{1}{1-x}$$

H : success $\pi \Rightarrow x = 1$

T : failure $1-\pi \Rightarrow x = 0$

H T H H H H T . . . . T

$x_1\ x_2\ x_3\ x_4\ x_5\ x_6\ x_7$ - - - - - - x_{100}

1 0 1 1 1 1 0 - - - - - . 0

$\left.\begin{matrix}\text{70 Heads}\\\text{30 Tails}\end{matrix}\right\}$

$$p(x_i \mid \pi) = \pi^{x_i} \cdot (1-\pi)^{(1-x_i)}$$

$$\left.\begin{matrix} Pr(x_i = 1 \mid \pi) = \pi^{1} \cdot (1-\pi)^{1-1} = \pi \\ Pr(x_i = 0 \mid \pi) = \pi^{0} \cdot (1-\pi)^{1-0} = 1-\pi \end{matrix}\right\} L(\pi \mid x) = \prod_{i=1}^{N} \left[\pi^{x_i} \cdot (1-\pi)^{1-x_i} \right]$$

$$\log L(\pi \mid x) = \sum_{i=1}^{N} \left[\overset{\checkmark}{x_i} \log(\overset{?}{\pi}) + (1-\overset{\checkmark}{x_i}) \log(1-\overset{?}{\pi}) \right] \Rightarrow \pi^{*} = ?$$

$$\frac{\partial \log L(\pi \mid x)}{\partial \pi} = \sum_{i=1}^{N} \left[x_i \cdot \frac{1}{\pi} + (1-x_i) \left(\frac{-1}{1-\pi} \right) \right] = 0$$

$$\pi^{*} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$\sum_{i=1}^{N} x_i$ → # of heads

N → # of tosses

Gaussian Density: $\mathcal{X} = \{x_i\}_{i=1}^{N}$

$$x_i \sim N(x_i; \mu, \sigma^2) \Rightarrow \mu^* = ? \qquad \sigma^{2^*} = ?$$

$$\sim \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \qquad -\infty < x_i < +\infty$$

$$\log \text{likelihood} = \log \prod_{i=1}^{N} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]\right]$$

$$\text{loglikelihood} = \sum_{i=1}^{N} \left[\left[-\frac{1}{2} \cdot \log(2\pi\overset{?}{\sigma}{}^2)\right] + \left[-\frac{(x_i-\overset{?}{\mu})^2}{2\underset{?}{\sigma}{}^2}\right]\right]$$

$$\left.\begin{aligned} \frac{\partial\, \text{loglikelihood}}{\partial \mu} &= 0 \\ \frac{\partial\, \text{loglikelihood}}{\partial \sigma^2} &= 0 \end{aligned}\right\} \text{Exercise} \Rightarrow \quad \mu^* = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \sigma^{2^*} = \frac{\sum_{i=1}^{N} (x_i - \mu^*)^2}{N}$$

Parametric Classification

Input: A training data set $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$

Output: A classifier

predicted class label

$$\hat{y}_{N+1} = \arg\max_{c} \; g_c(x_{N+1})$$

x_{N+1} → test data point

g_c → score function

$$\Pr(y=c \mid x) = \frac{p(x \mid y=c)\Pr(y=c)}{p(x)}$$

$p(x)$ → independent of class labels

$$\propto p(x \mid y=c)\Pr(y=c)$$

$\propto$ → proportional to

$$\underbrace{\log \Pr(y=c \mid x)}_{g_c(x)} \overset{+}{=} \log[p(x \mid y=c)] + \log[\Pr(y=c)] - \log[p(x)]$$

$\overset{+}{=}$ → equal up to a constant

$\log[p(x \mid y=c)]$ → class conditional densities are Gaussians (Normals)

$\log[\Pr(y=c)]$ → can be estimated using frequencies

$\log[p(x)]$ → constant

$$g_c(x) = \log\left[p(x \mid y=c)\right] + \log\left[\underbrace{Pr(y=c)}_{?}\right]$$

$$\mathcal{N}(x; \overset{?}{\mu_c}, \overset{?}{\sigma_c^2})$$

$$= \log\left[\underbrace{\frac{1}{\sqrt{2\pi\sigma_c^2}} \cdot \exp\left[-\frac{(x-\mu_c)^2}{2\sigma_c^2}\right]}\right] + \log\left[Pr(y=c)\right]$$

$$\mu_c^* = ? \qquad \sigma_c^{2*} = ?$$

$$\underbrace{Pr(y=c)}_{?}$$

$$\mu_c^* = \frac{\sum_{i=1}^{N}\left[x_i \cdot 1(y_i=c)\right]}{\sum_{i=1}^{N}\left[1(y_i=c)\right]}$$

$$\frac{N_c}{N} = \frac{\sum_{i=1}^{N} 1(y_i=c)}{N}$$

$$\sigma_c^{2*} = \frac{\sum_{i=1}^{N}\left[(x_i-\mu_c^*)^2\, 1(y_i=c)\right]}{\sum_{i=1}^{N}\left[1(y_i=c)\right]}$$

one function

$$1(\cdot) = \begin{cases} 1 & \text{if } \cdot \text{ is TRUE} \\ 0 & \text{otherwise} \end{cases}$$