Parametric Methods

Density Estimation

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

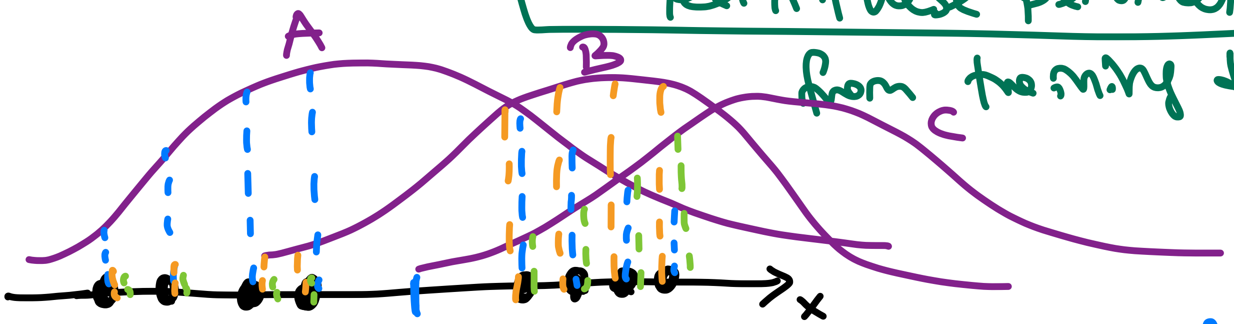
N data points

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

$\Downarrow$

parameters (?)

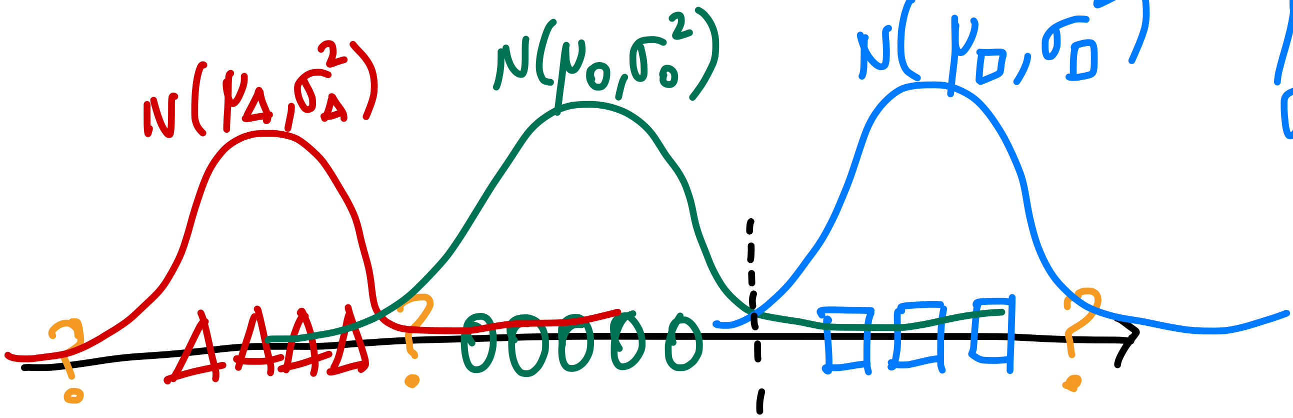
learn these parameters from training data



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

μ^* : the best μ parameter

σ^{2*} : the best σ^2 parameter



K=3 (# of classes)

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

△ ○ □

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

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

BAYES RULE $\Rightarrow$ $$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)}$$

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

?

x_{N+1}

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

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

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

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

pick the maximum one

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

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

↳ unknown parameters of p.

x_i's are i.i.d.

↳ identically & independently distributed

Likelihood $\equiv p(x_1, x_2, \ldots, x_N | \theta)$ → full joint

$\boxed{\begin{aligned} &P(A,B) \\ &= P(A)P(B) \end{aligned}}$

$$L(\theta | X) \equiv p(x_1|\theta)\, p(x_2|\theta) \ldots p(x_N|\theta)$$

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

$\boxed{\begin{aligned} &\log(abc) = \\ &\log(a) + \log(b) + \log c \end{aligned}}$

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

$\boxed{\begin{aligned} &\log(a^b) \\ &= b.\log(a) \end{aligned}}$

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

Bernoulli Density

$0 < \pi < 1 \Rightarrow$ success probability

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

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

$\boxed{\dfrac{\partial \log(x)}{\partial x} = \dfrac{1}{x}}$

$\boxed{\dfrac{\partial \log(1-x)}{\partial x} = -\dfrac{1}{1-x}}$

coin $\Rightarrow$ H T H H H H T . . . . T

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

1 0 1 1 1 1 0 0

} 70 heads, 30 tails

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

$$Pr(x_i = 1|\pi) = \pi^{1} \cdot (1-\pi)^{1-1} = \pi$$

$$Pr(x_i = 0|\pi) = \pi^{0} \cdot (1-\pi)^{1-0} = 1-\pi$$

$$\mathcal{L}(\pi|\mathcal{X}) = \prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i}$$

$$\log \mathcal{L}(\pi|\mathcal{X}) = \sum_{i=1}^{N}\left[x_i \cdot \log(\pi) + (1-x_i)\log(1-\pi)\right] \Rightarrow \pi^* = ?$$

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

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

Gaussian Density

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

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

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

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

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

$$\frac{\partial \log \text{likelihood}}{\partial \mu} = 0 \quad \& \quad \frac{\partial \log \text{likelihood}}{\partial \sigma^2} = 0$$

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

$$\sigma^{2*} = \frac{\sum_{i=1}^{N}(x_i-\mu^*)^2}{N}$$

Parametric classification:

Input: a training data set

$$\mathcal{X} = \{(\underbrace{x_i}_{\text{input}}, \overbrace{y_i}^{\text{output}})\}_{i=1}^{N}$$

Output: A classifier

$$\hookrightarrow \hat{y}_{N+1} = \arg\max_c \; \underbrace{g_c}_{\text{score function for class \#c}}(\overbrace{x_{N+1}}^{\text{test data point}})$$

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

$\rightarrow$ independent of class labels

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

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

$\hookrightarrow$ equal up to a constant

$$g_c(x) = \log\left[\underbrace{p(x \mid y=c)}_{N(x;\,\mu_c,\,\sigma_c^2)}\right] + \log\left[\underbrace{\Pr(y=c)}_{\text{frequency of class \# c in our training set}}\right]$$

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

$\mu_c^* = ?$ $\quad \sigma_c^{2*} = ?$ $\quad \hat{\Pr}(y=c) = ?$

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

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

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

one function

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

MODEL PARAMETERS

$K = \#$ of classes

$$\mu_1^*, \mu_2^*, \ldots\ldots, \mu_K^* \Rightarrow K$$

$$\sigma_1^{2*}, \sigma_2^{2*}, \ldots\ldots, \sigma_K^{2*} \Rightarrow K$$

$$\hat{\Pr}(y=1), \hat{\Pr}(y=2), \ldots\ldots, \hat{\Pr}(y=K) \Rightarrow K-1$$

total # of parameters $= 3K - 1$