Density Estimation!

$\mathcal{X} = \{x_i\}_{i=1}^{N}$ N samples (data points)

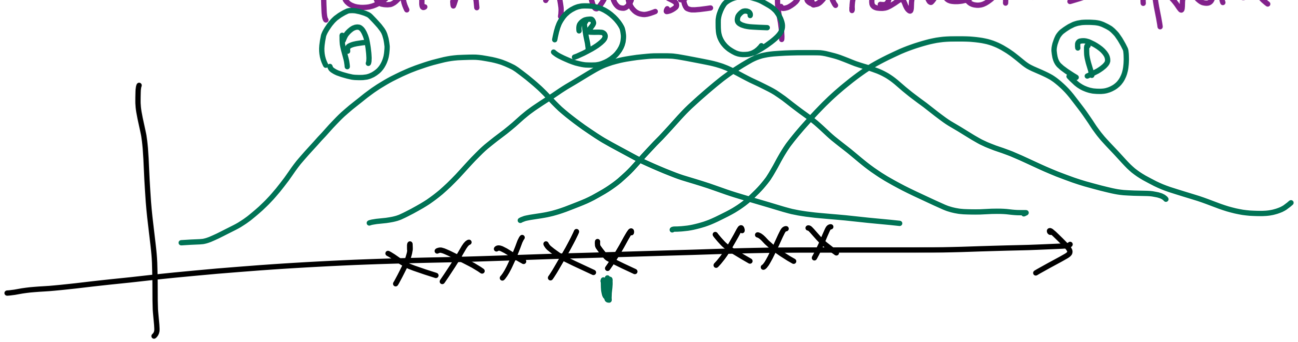
$x_i \sim \underbrace{p(x)}_{\text{unknown}} \quad \forall i \Rightarrow$ probability distribution

$\Downarrow$

parameters (?)

DENSITY ESTIMATION

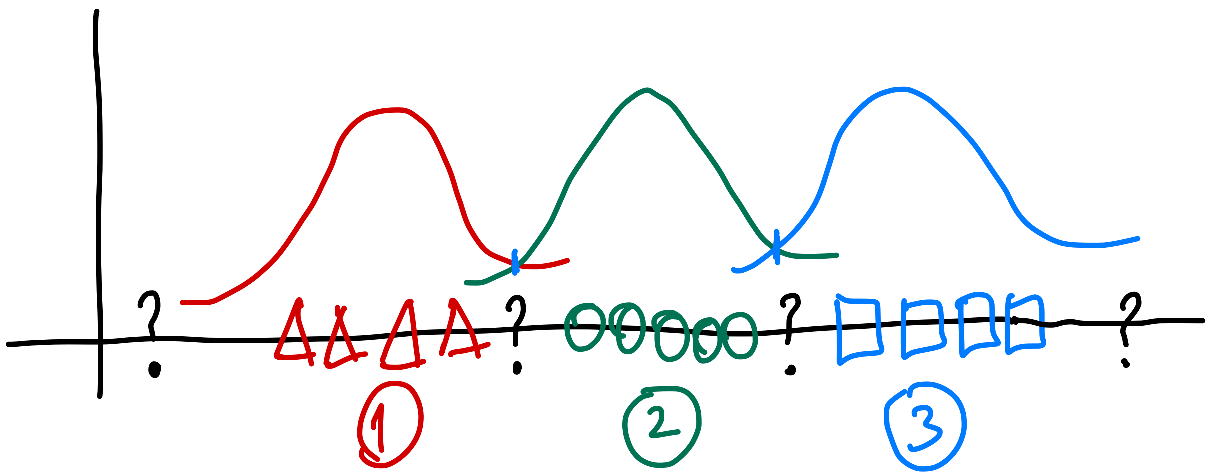
learn these parameters from training data

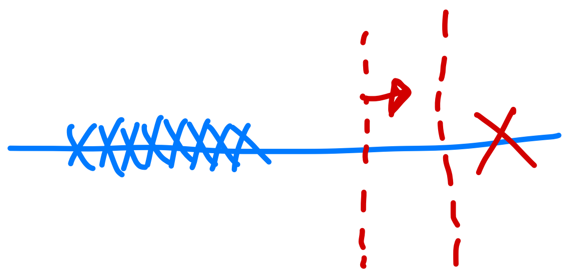


$x_i \sim N(x; \mu, \sigma^2)$

μ^*: the best μ parameter

σ^{2*}: the best σ^2 parameter





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

class densities $\Rightarrow p(x \mid y=c)$ ⤳ density estimation

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

BAYES RULE

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

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

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

a new data point $\underbrace{x_{N+1}}$ $\Rightarrow P(y=c \mid x_{N+1})$

$$\left.\begin{array}{l} \rightarrow P(y=1 \mid x_{N+1}) \\ \rightarrow P(y=2 \mid x_{N+1}) \\ \rightarrow P(y=3 \mid x_{N+1}) \end{array}\right\} \text{pick the maximum}$$

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

$$X = \{x_i\}_{i=1}^{N} \qquad x_i \sim p(x \mid \theta)$$

$\hookrightarrow$ unknown parameters

$\log(a^b) = b.\log(a)$

x_i's are i.i.d.

identically & independently distributed

$\log(a.b.c) = \log(a) + \log(b) + \log(c)$

$$\text{Likelihood} \equiv p(x_1, x_2, \ldots, x_N \mid \theta)$$

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

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

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

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

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

Bernoulli density: $0 < p < 1$ $\hookrightarrow$ success probability

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

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

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

70 heads / 30 tails

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

$\Rightarrow$ 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

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

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

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

$$P(x_i = 0 \mid p) = p^0 (1-p)^{1-0} = 1-p$$

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

$$\frac{\partial \log L(p \mid \mathcal{X})}{\partial p} = \sum_{i=1}^{N} \left[x_i \cdot \frac{1}{p} - (1-x_i) \frac{1}{1-p} \right] = 0 \Rightarrow p^{*} = \frac{\sum_{i=1}^{N} x_i}{N}$$

# of heads

Gaussian Density:

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

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

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

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

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

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

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

Parametric Classification:

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

$\hookrightarrow y_i \in \{1, 2, \ldots, K\}$

Output: A classifier

test (unseen) data point

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

$\hookrightarrow$ score function for class #c

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

$p(x) \Rightarrow$ independent of class labels

$$P(y=c \mid x) \propto p(x \mid y=c)\,P(y=c)$$

$\hookrightarrow$ "proportional to"

$$\log P(y=c \mid x) = \log(p(x \mid y=c)) + \log(P(y=c)) - \cancel{\log(p(x))}$$

constant

$$\stackrel{+}{=} \log(p(x \mid y=c)) + \log(P(y=c))$$

$\hookrightarrow$ "equal up to a constant"

$$g_c(x) = \log\big(\underbrace{p(x\mid y=c)}_{\mathcal{N}(x;\,\mu_c,\,\sigma_c^2)}\big) + \log\big(\underbrace{P(y=c)}_{\text{frequency of class } \#c \text{ in our data set.}}\big)$$

$$= \log\Bigg[\underbrace{\frac{1}{\sqrt{2\pi\sigma_c^2}} \cdot \exp\left[-\frac{(x-\mu_c)^2}{2\sigma_c^2}\right]}_{\mu_c^* = ?\quad \sigma_c^{2*} = ?}\Bigg] + \log\big(\underbrace{P(y=c)}_{\frac{N_c}{N} = \frac{\sum_{i=1}^{N} 1(y_i=c)}{N}}\big)$$

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

↳ # of data points that belong to class # c.

"one" function

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

$\mu_1^*, \mu_2^*, \ldots, \mu_K^* \quad \} \; K$

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

$\hat{p}(y=1), \hat{p}(y=2), \ldots, \hat{p}(y=K) \quad \} \; K-1$

total # of parameters $= 3K-1$.