Density Estimation

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

N data points

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

$\Downarrow$

parameters (?)

DENSITY ESTIMATION

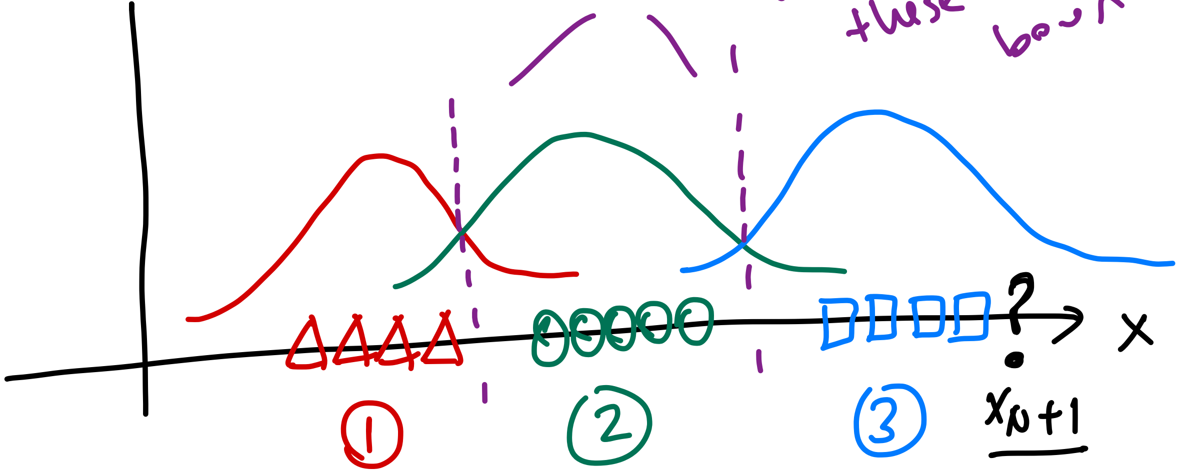
learning these parameters from training data

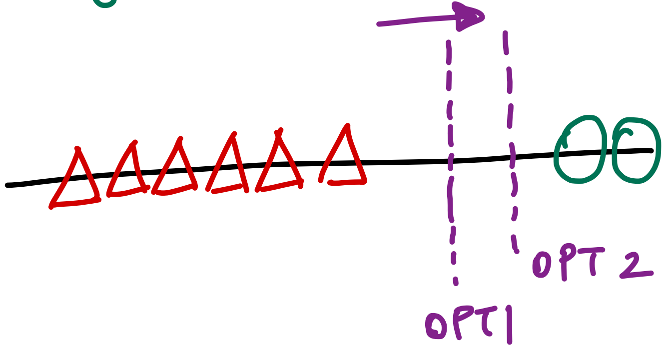
$x_i \sim \mathcal{N}(x_i; \mu, \sigma^2)$

μ^* : the best μ parameter

σ^{2*} : the best σ^2 parameter

if frequencies were equal these would be boundaries





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

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

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

BAYES RULE:

$$P(B \mid A) = \frac{\overbrace{P(A, B)}^{\text{joint distribution}}}{P(A)} = \frac{P(A \mid B) \cdot P(B)}{P(A)}$$

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

1003 $\qquad$ 3

$\boxed{1004}$ $\qquad$ $\boxed{4}$

1002 $\qquad$ 2

test data point

$\Downarrow$

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

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

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

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

x_i's are i.i.d.

↳ identically & independently distributed

Likelihood $\equiv p(x_1, x_2, x_3, \ldots, x_N | \theta)$ ⇒ full joint

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

$$\boxed{\equiv \prod_{i=1}^{N} p(x_i|\theta)} \Rightarrow \theta^* = \arg\max_{\theta} L(\theta|\mathcal{X})$$

θ^*: best θ

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

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

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

Bernoulli Density:

$0 < \pi < 1$

$\hookrightarrow$ success probability

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

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

$$\left.\begin{array}{ll}(H)\ \text{success} : & \pi \\ (T)\ \text{failure} : & 1-\pi\end{array}\right] \begin{array}{l}\Rightarrow x = 1 \\ \Rightarrow x = 0\end{array}$$

A Head / Tail $= (1/2, 1/2)$

B Head / Tail $= (7/10, 3/10)$

coin $\Rightarrow$

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

$\Big\}$ 70 heads, 30 tails

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

$$\left\{\begin{array}{l} P(x_i=1|\pi) = \pi^1 \cdot (1-\pi)^{1-1} = \pi \\ P(x_i=0|\pi) = \pi^0 \cdot (1-\pi)^{1-0} = 1-\pi \end{array}\right\}$$

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

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

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

# of heads

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

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

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

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

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

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

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

sample mean

sample variance

Parametric Classification:

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

Output: A classifier

test data point

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

score function for class # c.

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

$\Rightarrow$ independent of class labels.

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

"proportional to"

constant

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

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

"equal up to a constant"

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

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

N_c → # of data points that belong to class #c.

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

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

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

$$\begin{bmatrix} \mu_1^*, \mu_2^*, \dots\dots, \mu_K^* \\ \sigma_1^{2*}, \sigma_2^{2*}, \dots\dots, \sigma_K^{2*} \\ \hat{P}(y=1), \hat{P}(y=2), \dots, \hat{P}(y=K) \end{bmatrix} \begin{array}{l} \Rightarrow K \\ \Rightarrow K \\ \Rightarrow K-1 \end{array}$$

$+$

total # of parameters $= 3K - 1$

$H = 1/2 \quad T = 1/2$

$H = 3/4 \quad T = 1/4$

$H\,H\,H\,T \rightarrow$ likelihood $\Rightarrow \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{8}{64}$

$H\,H\,H\,T \rightarrow$ likelihood $\Rightarrow \frac{3}{4}\frac{3}{4}\frac{3}{4}\frac{1}{4} = \frac{27}{64}$