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

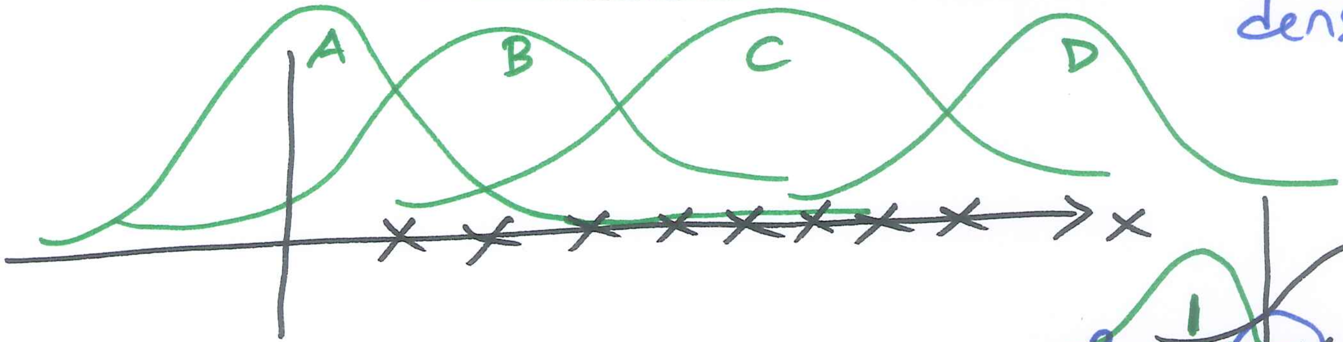
$x_i \sim p(x)$ ⇒ probability distribution
↳ unknown

⇓

parameter(s)

learn them from training data

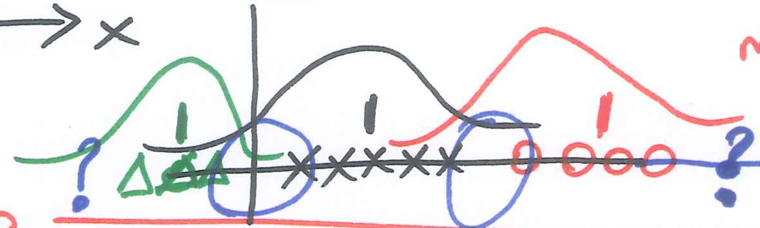
density estimation



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

mean ? μ^*

variance ? σ^{2*}



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

class densities ⇒ $p(x \mid y=c)$

prior distribution ⇒ $p(y=c)$

BAYES RULE

POSTERIOR

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

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

$p(y=1 \mid x)$, $p(y=2 \mid x)$, $p(y=3 \mid x)$

MAXIMUM LIKELIHOOD ESTIMATION (MLE)

$\mathcal{X} = \underbrace{\{x_i\}_{i=1}^{N}}_{}$ $\quad x_i \sim p(x \mid \theta)$ $\hookrightarrow$ (unknown) parameters

x_i's are i.i.d. $\qquad P(A, B, C) = P(A)P(B)P(C)$

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

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

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

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

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

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

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

$$\boxed{N(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]}$$

Bernoulli Density:

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

$P(x = \text{head}) = 1/2$

$P(x = \text{tails}) = 1/2$

success: $P \Rightarrow x = 1$ $\quad 0 < p < 1$

failure: $1-P \Rightarrow x = 0$

$$\boxed{\begin{aligned} p(x) &= p^x (1-p)^{1-x} \\ E[x] &= \sum x p(x) \\ &= 0.(1-p) + 1.p \\ &= p \\ VAR[x] &= E[x^2] - E[x]^2 \\ &= p(1-p) \end{aligned}}$$

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

x_1 x_2 . . . . . . x_{100}

1 0 0 1 1 1 0

$p(x|p)$

$$\hookrightarrow ? \log L(p|\mathcal{X}) = \log \prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}$$

$$\frac{\partial \log x}{\partial x} = \frac{1}{x} \qquad \text{Loglikelihood} = \sum_{i=1}^{N} \log\left[p^{x_i} . (1-p)^{1-x_i}\right]$$

$$\frac{\partial \log(1-x)}{\partial x} = (-1).\frac{1}{1-x} \qquad = \sum_{i=1}^{N} \left(x_i \log(\underline{p}) + (1-x_i)\log(1-\underline{p})\right)$$

$\ast$ $P = ?$

$$\frac{\partial \text{Loglikelihood}}{\partial p} = \sum_{i=1}^{N} \left[x_i \frac{1}{p} + (1-x_i).\frac{(-1)}{1-p}\right] = 0$$

$$\ast \quad \hat{p} = \left[\frac{\sum_{i=1}^{N} x_i}{N}\right]$$

Gaussian density: $\mathcal{X} = \{x_1, \ldots, x_N\} = \{x_i\}_{i=1}^{N}$

$x_i \sim N(x; \mu, \sigma^2)$ $\quad \mu^* = ?$ $\quad \sigma^{2*} = ?$ $\quad \frac{1}{\sqrt{2\pi\sigma^2}} = (2\pi\sigma^2)^{-1/2}$

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

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

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

$$\frac{\partial \mathcal{L}(\mu, \sigma^2 | \mathcal{X})}{\partial \mu} = 0 \qquad \frac{\partial \mathcal{L}(\mu, \sigma^2 | \mathcal{X})}{\partial \sigma^2} = 0$$

$$\boxed{\mu^* = \frac{\sum_{i=1}^{N} x_i}{N}} \qquad \boxed{\sigma^{2*} = \frac{\left[\sum (x_i - \mu^*)^2\right]}{N}}$$

Sample mean $\qquad$ Sample variance.

Parametric Classification

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

$\hookrightarrow$ K classes $\Rightarrow y_i \in 1 \ldots K$

Output: a classifier

$$\begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_K(x) \end{bmatrix} \qquad y^* = \arg\max_c \, g_c(x^*)$$

$\hookrightarrow$ score function for class $\underline{\underline{c}}$

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

$p(x)$ } independent of class labels

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

$\hookrightarrow$ $\propto$: "is proportional to"

$$\log p(y=c \mid x) = \log p(x \mid y=c) + \log(y=c) - \log p(x)$$

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

$\hookrightarrow$ $\stackrel{+}{=}$ "equal up to a constant"

$$g_c(x) = \log p(y=c \mid x) \stackrel{+}{=} \underbrace{\log p(x \mid y=c)}_{\mu_c,\ \sigma_c^2} + \underbrace{\log p(y=c)}_{\widehat{p(y=c)}}$$

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

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

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

$$\widehat{p(y=c)} = \qquad \widehat{p(y=c)} = \frac{\overbrace{\sum_{i=1}^{N} 1(y_i=c)}^{N_c}}{N} = \frac{N_c}{N}$$

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

$$\left.\begin{array}{ll} \mu_1^*, \mu_2^*, \ldots, \mu_K^* & \} K \checkmark \\ \sigma_1^{2*}, \sigma_2^{2*}, \ldots, \sigma_K^{2*} & \} K \checkmark \\ \widehat{p(y=1)}, \widehat{p(y=2)}, \ldots, \widehat{p(y=K)} & \} K-1 \checkmark \end{array}\right\} \text{total \# of parameters } [3*K-1]$$