$\boxed{X|y=0}$ $\boxed{X|y=1}$

$\boxed{y}$

Testing

For a $X^*$

$$P(Y=1|X=x^*) = \frac{P(X=x^*|Y=1)\,P(Y=1)}{P(X=x^*|Y=0)\,P(Y=0) + P(X=x^*|Y=1)\,P(Y=1)}$$

$$P(Y=0|X=x^*) = \frac{\qquad}{\qquad}$$

Choosing the form of $P(X=x^*|Y=1 \text{ or } 0)$ and $P(Y)$ is zeroth step.

If $X \in [0,1]^d$

What form should $x$ take?

| Assume $d=1$ | $d=2$ | $d=3$ | $d=d$ |
|---|---|---|---|
| $P(X=0)$<br>$P(X=1)$ | $P(X_1=0, X_2=0)$<br>$P(X_1=1, X_2=0)$<br>$P(X_1=0, X_2=1)$<br>$P(X_1=1, X_2=1)$ | ⋮ | |
| 1 parameter | 3 parameter. | 7 parameter | $(2^d-1)$ |
| For two classes $-\ 2\times1$ | $2\times3$ | $2\times7$ | $2\times(2^d-1)$ |

Problem: More parameters $\Rightarrow$ More data!

$d=2$

Assume that $X_1$ and $X_2$ are independent.

$$P(X_1=0, X_2=0) = P(X_1=0)\,P(X_2=0)$$

$$P(X_1=0, X_2=0) = P(X_1=0) * P(X_2=0)$$

$$P(X_1=1, X_2=0) = P(X_1=1) * P(X_2=0)$$

$$P(X_1=0, X_2=1) = P(X_1=0) * P(X_2=1)$$

$$P(X_1=1, X_2=1) = P(X_1=1) * P(X_2=1)$$

All I need is $P(X_1=1)$ and $P(X_2=1)$

---

Naive Bayes Classifier. (NBC)

$$P(X=x^* \mid y=1) = \prod_{j=1}^{d} p(x_j^* \mid y=1)$$

$x_j^*$ is the $j^{th}$ feature in $x^*$

## NBC

$\theta$ → for Y

$P(Y=1) = \theta$

$P(Y=0) = 1-\theta$

$\theta_{11}, \theta_{12}, \theta_{13} \cdots \theta_{1D}$

$\theta_{21}, \theta_{22}, \theta_{23} \cdots \theta_{2D}$

$P(X_1 = x_1 \mid y=1)$ ← $\theta_{11}, \theta_{12}, \theta_{13} \cdots \theta_{1D}$

$x_1 \rightarrow (0,1)$

$P(X_2 = x_2 \mid y=0)$ ← $\theta_{21}, \theta_{22}, \theta_{23} \cdots \theta_{2D}$

$\theta_{11} : \quad P(X_1 = 1 \mid y = 1) \quad : \quad P(X_1 = 0 \mid y=1) = 1 - \theta_{11}$

$\theta_{21} : \quad P(X_1 = 1 \mid y = 0)$

---

Is $\theta_{21} = 1 - \theta_{11}$ ? NO

---

MLE:

$$D = \begin{bmatrix} x_1, & y_1 \\ x_2, & y_2 \\ x_3, & y_3 \\ \vdots & \vdots \\ x_N & y_N \end{bmatrix}$$

Training data

Likelihood

$$L(D \mid \theta) = \prod_{i=1}^{N} P(X_i = x_i, Y_i = y_i)$$

$$\log L(D \mid \theta) = \sum_{i=1}^{N} \log P(X_i = x_i, Y_i = y_i)$$

What is $P(X_i = x_i, Y_i = y_i)$

$$= P(Y_i = y_i)\, P(\underline{X_i = x_i} \mid Y_i = y_i) \longrightarrow \text{independent}$$

$$= P(Y_i = y_i) \prod_{j=1}^{D} P(X_{ij} = x_{ij} \mid Y_i = y_i)$$

$$= \theta^{y_i}(1-\theta)^{1-y_i}\left[\prod_{j=1}^{D} \theta_{y_i j}^{x_{ij}} (1-\theta_{y_i j})^{1-x_{ij}}\right]$$

---

$x_i \rightarrow \boxed{x_{i1} \mid x_{i2} \mid \cdots \mid x_{iD}}$

Assume $i = 1$ and $j = 1$

$P(X_{11} = x_{11} \mid Y_1 = y_1)$

$\hookrightarrow \theta_{11}$

$$ll(D \mid \theta)$$

$$= \sum_{i=1}^{N}\left[\log\theta^{y_i} + \log(1-\theta)^{(1-y_i)} + \sum_{j=1}^{D}\left\{\log \theta_{y_i j}^{x_{ij}} + \log(1-\theta_{y_i j})^{1-x_{ij}}\right]\right]$$

$$= \sum_{i=1}^{N}\left[\left(y_i \log\theta + (1-y_i)\log(1-\theta)\right) + \sum_{j=1}^{D}\left[x_{ij}\log\theta_{y_i j} + (1-x_{ij})\log(1-\theta_{y_i j})\right]\right]$$

---

$\theta, [\theta_{y_i j})$

If we compute $\frac{\partial}{\partial \theta} ll(D|\theta)$

$$\frac{\partial}{\partial \theta} ll(D|\theta) = \sum_{i=1}^{N} \frac{\partial}{\partial \theta}\left[y_i \log \theta + (1-y_i)\log(1-\theta)\right] + 0$$

$$= \sum_{i=1}^{N}\left[\frac{y_i}{\theta} - \frac{1-y_i}{1-\theta}\right]$$

Setting $\frac{\partial}{\partial \theta} ll(D|\theta) = 0$

$$\boxed{\theta = \frac{N_1}{N}}$$

$N_1$ = # times $y_i$ is 1 in the data set.

$$\boxed{\theta_{1j} = \frac{N_{1j}}{N_1}}$$

$N_{1j}$ - # times $x_{ij}$ is 1 when $y_i$ is 1

$$\boxed{\theta_{0j} = \frac{N_{0j}}{N_0}}$$

# Naive Bayes Example

| $X_1$ | $X_2$ | $X_3$ | $X_4$ | $Y$ |
|---|---|---|---|---|
| 1 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 1 |
| 1 | 0 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 0 | 0 |
| 1 | 1 | 0 | 1 | 1 |

~~$P(1, 0, 1, 1)$~~

$P(Y=1 \mid X = 1,1,1,1) \hookrightarrow ?$

$$P(Y=1 \mid X_1=1, X_2=1, X_3=1, X_4=1)$$

$$= \frac{P(X_1=1, X_2=1, X_3=1, X_4=1 \mid Y=1)\, P(Y=1)}{P(X_1=1, X_2=1, X_3=1, X_4=1 \mid Y=1)\, P(Y=1) + P(X_1=1, X_2=1, X_3=1, X_4=1 \mid Y=0)\, P(Y=0)}$$

$\theta = \frac{5}{8}$ $\qquad \theta = \frac{N_1}{N}$

| $y=0$ | $y=1$ |
|---|---|
| $\theta_{01} = \frac{1}{3}$ | $\theta_{11} = \frac{3}{5}$ |
| $\theta_{02} = \frac{1}{3}$ | $\theta_{12} = \frac{3}{5}$ |
| $\theta_{03} = \frac{1}{3}$ | $\theta_{13} = \frac{2}{5}$ |
| $\theta_{04} = \frac{1}{3}$ | $\theta_{14} = \frac{4}{5}$ |

First compute:

$P(X_1=1, X_2=1, X_3=1, X_4=1 \mid Y=1)$

$= P(X_1=1 \mid Y=1)\, P(X_2=1 \mid Y=1)\, P(X_3=1 \mid Y=1)\, P(X_4=1 \mid Y=1)$

$= \theta_{11} * \theta_{21} * \theta_{31} * \theta_{41}$

$P(X_2=1 \mid Y=1) = \theta_{21}$

$P(X_2=0 \mid Y=1) = (1-\theta_{21})$

$P(X_1=1, X_2=1, X_3=1, X_4=1 \mid Y=0)$

$= \theta_{01} * \theta_{02} * \theta_{03} * \theta_{04}$

$P(Y=1) = \theta = 5/8$

$P(Y=0) = 1-\theta = 3/8$

If $X_4$ takes three values $\rightarrow (0, 1, 2)$

$\theta_{020} \quad \theta_{021} \quad \theta_{022}$

$\theta_{120} \quad \theta_{121} \quad \theta_{122}$

Assumption : Data is generated by a Gaussian Dis

$(\mu, \sigma) \longrightarrow ?$

0.8
0.7
0.3 $\longrightarrow$ Learn $\mu, \sigma$
1.2
1.1

~~pdf of~~ $p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$

Training Data

lik. $= p(0.8) * p(0.7) * p(0.3) * p(1.2) * p(1.1)$

loglik $= \sum_{i=1}^{5} \log p(x_i)$

$L(\mu, \sigma)$

$\hat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} x_i$ [Sample mean]

$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum (x_i - \mu)^2$ [Sample variance]

What happens if $x \in \mathbb{R}^2$

$x = [0.7, 0.3]$

$$\boxed{\begin{array}{l} P(x=0.7) \times \\ p(0.7) \end{array}}$$

Naive Bayes:

$$P(\mathbf{X} \mid y = \text{malignant}) = P(x_1 \mid y=m)\, p(x_2 \mid y=m)\, P(x_3 \mid y=m)$$

| $x_1$ | $x_2$ | $x_3$ |
|---|---|---|

cont or ord, binary

What happens if we have

| 0.7 | 0.3 | 0.4 |
|---|---|---|

$$P(x_1 = 0.7 \mid y=m)\, p(x_2 = 0.3 \mid y=m)\, p(x_3 = 0.4 \mid y=m)$$

We can model each $x_1, x_2, x_3$ as univariate Gaussian random variable.

$$\Rightarrow \quad P(x_1 = 0.7 \mid y=m) = \frac{1}{\sqrt{2\pi}\,\sigma_{1m}} \exp\left[-\frac{1}{2\sigma_{1m}^2}(0.7 - \mu_{1m})^2\right]$$

---

$$p(y \mid x) \propto p(y) \prod_j p(x_j \mid y)$$

$$= p(y) \prod_{j=1}^{D} \left( \frac{1}{\sqrt{2\pi\sigma_j^2}} \cdot \exp - \frac{(x_j - \mu_j)^2}{2\sigma_j^2} \right)$$

$$= p(y) \frac{1}{(2\pi)^{D/2} * (\sigma_1^2 * \sigma_2^2 \cdots * \sigma_D^2)^{1/2}} \exp\left[-\frac{1}{2}\sum \frac{(x_j - \mu_j)^2}{\sigma_j^2}\right]$$

$$\text{let } \Sigma = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_D^2 \end{bmatrix}$$

$$(\sigma_1^2 * \sigma_2^2 * \cdots * \sigma_D^2)^{1/2} = |\Sigma|^{1/2}$$

$$\boxed{\sum_{j=1}^{D} \left(\frac{x_j - \mu_j}{\sigma_j^2}\right)^2 = (x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$\text{where } \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_D \end{bmatrix}$$

$$\Rightarrow p(y|x) = p(y) \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

$\Sigma$ - diagonal matrix

$$\Sigma^{-1} = \begin{bmatrix} 1/\sigma_1^2 & & & 0 \\ & 1/\sigma_2^2 & & \\ & & \ddots & \\ 0 & & & 1/\sigma_D^2 \end{bmatrix}$$

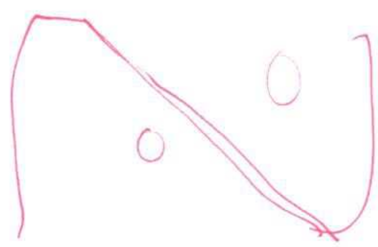
$$p(y|\mathbf{x}) = \frac{p(y)\, p(\mathbf{x}|y)}{\text{Den.}}$$

vector

where $\mathbf{X}$ is modeled as a multivariate Gaussian

---

$\Sigma$ need not be a diagonal matrix.



## Quadratic Discriminant Analysis.

QDA:

$$p(y=1|x) = \frac{p(y=1)\, p(x|y=1)}{p(y=1)\, p(x|y=1) + p(y=2)\, p(x|y=2)}$$

where
$$p(x|y=1) = \frac{1}{(2\pi)^{D/2} |\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1} (x-\mu_1)\right]$$

$$p(x|y=2) = \frac{1}{(2\pi)^{D/2} |\Sigma_2|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_2) \Sigma_2^{-1} (x-\mu_2)\right]$$

# How to train your QDA?

Training Data:

| | |
|---|---|
| $X_1$ | $y_1$ |
| $X_2$ | $y_2$ |
| $\vdots$ | $\vdots$ |
| $X_N$ | $y_N$ |

Split data into two parts.

(1) $X_1, X_2, \ldots, X_M \rightarrow y_i$ is 1 $\rightarrow$ MLE to get $\mu_1\ \Sigma_1$

(2) $X_{M+1}, \ldots, X_N \rightarrow y_i$ is 2 $\rightarrow$ MLE to get $\mu_2\ \Sigma_2$

(3) Compute $\Theta$ for $p(Y)$

---

## Linear Discriminant Analysis (LDA)

In LDA: $\Sigma_1 = \Sigma_2 = \Sigma$

In training: $\mu_1$ and $\mu_2$ and $\Theta$ are estimated in the same way as QDA

For $\Sigma$: We use the entire training data and do MLE for $\Sigma$

# LDA:

$$p(y=1|x) \propto p(y=1)\, p(x|y=1)$$

$$= p(y=1) \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right]$$

$$p(y=2|x) = p(y=2) \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_2)\Sigma^{-1}(x-\mu_2)\right]$$

Let us assume that $p(y=1) = p(y=2)$

$$\Rightarrow p(y=1|x) \propto \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right]$$

$$p(y=2|x) \propto \exp\left[-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right]$$

---

Just compute $\underline{(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)} \longrightarrow D_1$

and $(x-\mu_2)^T\Sigma^{-1}(x-\mu_2) \longrightarrow D_2$

If $\underline{D_1 < D_2}$

$$\Rightarrow \exp\left(-\frac{1}{2}D_1\right) \geq \exp\left(-\frac{1}{2}D_2\right)$$

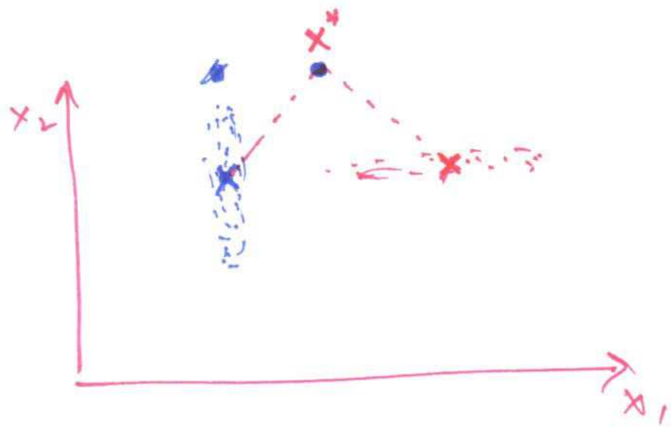
$$\Rightarrow p(y=1|x) \geq p(y=2|x)$$

---

What is $(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)$

Mahalanobis distance

$$\underline{(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)}$$




Euclidean Distance

$x, \mu_1 \in \mathbb{R}^D$

$$\sum_{i=1}^{D} (x_i - \mu_{1i})^2$$

$$= \underline{(x-\mu_1)^T (x-\mu_1)}$$