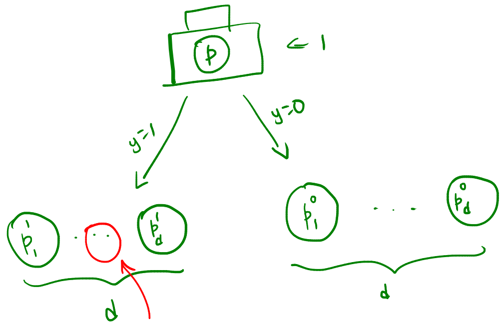


CLASS CONDITIONAL INDEPENDENCE

$2d+1$

PARAMETER ESTIMATION

$$p, \{p_1^1, \dots p_d^1\}, \{p_1^0, \dots, p_d^0\}$$

MAX. LIKELIHOOD ESTIMATES

$$1 \rightarrow \hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i \rightarrow \{\text{Fraction of spam emails in the dataset}\}$$

$\forall j \in \{1, \dots, d\}$
$\forall y \in \{0, 1\}$

$$\hat{p}_j^y = \frac{\sum_{i=1}^{n} \mathbb{1}(f_j^i = 1, y_i = y)}{\underbrace{\sum_{i=1}^{n} \mathbb{1}(y_i = y)}} \leftarrow \text{Number of emails with label } y.$$

$\hookrightarrow$ Fraction of y-labelled emails that contain the j^{th} word.

PREDICTION

Given $x^{test} \in \{0,1\}^d$, what is $\hat{y}^{test}$?

$$P(y^{test} = 1 / x^{test}) > P(y^{test} = 0 / x^{test})$$

$$\Rightarrow \hat{y}^{test} = 1$$

$$= 0 \quad \text{otherwise.}$$

How to obtain $P(y/x)$ from $P(y)$ and $P(x/y)$?

BAYES RULE !

$$P\left(y^{test}=1 \mid x^{test}\right) = \frac{P\left(x^{test} / y^{test}=1\right) \cdot P\left(y^{test}=1\right)}{P\left(x^{test}\right)}$$

$$P\left(y^{test}=0 \mid x^{test}\right) = \frac{P\left(x^{test} / y^{test}=0\right) \cdot P\left(y^{test}=0\right)}{P\left(x^{test}\right)}$$

$$\underbrace{P\left(x^{test} / y^{test}=1\right)} \cdot P\left(y^{test}=1\right)$$

$$= P\left(x^{test}=[f_1 \; f_2 \; \dots \; f_d] / y^{test}=1\right) \cdot P\left(y^{test}=1\right)$$

$$= \left(\prod_{j=1}^{d} \left(\hat{p}_j^{\,1}\right)^{f_j} \left(1-\hat{p}_j^{\,1}\right)^{(1-f_j)}\right) \cdot \hat{p}$$

$$X_{test} = \begin{array}{cccc} f_1 & f_2 & f_3 & f_4 \\ \downarrow & \downarrow & \downarrow & \downarrow \end{array} \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}$$

if

$$\left(\prod_{j=1}^{d} \left(\hat{p}_j^{\,1}\right)^{f_j} \left(1-\hat{p}_j^{\,1}\right)^{(1-f_j)}\right) \cdot \hat{p} > \left(\prod_{j=1}^{d} \left(\hat{p}_j^{\,0}\right)^{f_j} \left(1-\hat{p}_j^{\,0}\right)^{(1-f_j)}\right) \cdot (1-\hat{p})$$

$\Rightarrow$ PREDICT $\hat{y}^{test} = 1$

else $\hat{y}^{test} = 0$.

MODEL uses 2 main Things

NAIVE BAYES ALGORITHM.

- CLASS CONDITIONAL INDEPENDENCE
- $\boxed{\text{BAYES}}$ THEOREM

- may not hold in practice
- $\boxed{\text{NAIVE}}$ Assumption
- Still works well in practice.

PITFALLS IN NAIVE BAYES TO WATCH OUT FOR.

- IF a word does not appear in the train set but appears in a test data point,

$\hat{p}_j^1 = 0 \qquad \hat{p}_j^0 = 0$

$$P\left(y^{test}=1 \,\middle|\, x^{test}=[f_1, f_2 \dots \underset{\substack{\uparrow \\ \text{zebra} \\ =1}}{f_d}]\right) \propto \underbrace{\left(\prod_{i=1}^{d} \underbrace{(\hat{p}_i^1)}_{0}^{f_i} \; \underbrace{(1-\hat{p}_i^1)}_{1}^{(1-f_i)}\right)}_{0} \hat{p}$$

$$P\left(y^{test}=0 \,\middle|\, x^{test}=[f_1 \dots \underset{\substack{\uparrow \\ \text{zebra} \\ =1}}{f_d}]\right) \propto \underbrace{\left(\prod_{i=1}^{d} \underbrace{(\hat{p}_i^0)}_{0}^{f_i} \; \underbrace{(1-\hat{p}_i^0)}_{1}^{(1-f_i)}\right)}_{0} (1-\hat{p})$$

Possible Fix

- Can add two "pseudo" emails with all words present – one email has label 0 and another has label 1

LAPLACE SMOOTHING

$$\begin{array}{lc} x_1^{pseudo} & y_1^{pseudo} \\ [1 \; 1 \; 1 \; 1 \; 1 \; \dots \; 1] & 1 \\ [1 \; 1 \; 1 \; \dots \; 1] & 0 \\ x_0^{pseudo} & \end{array}$$

DECISION FUNCTION OF NAIVE BAYES.

Given x_{test}; $y_{test}=1$ if

$$\frac{P(y_{test}=1 / x_{test})}{P(y_{test}=0 / x_{test})} \geq 1$$

$$\log\left(\frac{P(y_{test}=1 \mid x_{test})}{P(y_{test}=0 \mid x_{test})}\right) \geq 0$$

$$\log\left(\frac{\underbrace{P(x_{test} \mid y_{test}=1)\cdot P(y_{test}=1)} \,/\, P(\cancel{x_{test}})}{\underbrace{P(x_{test} \mid y_{test}=0)\cdot P(y_{test}=0)} \,/\, P(\cancel{x_{test}})}\right) \geq 0$$

$x_{test} = [f_1\ f_2 \cdots f_d]$

$$\log\left(\prod_{i=1}^{d} \frac{(\hat{p}_i^1)^{f_i}\,(1-\hat{p}_i^1)^{(1-f_i)} \cdot \hat{p}}{(\hat{p}_i^0)^{f_i}\,(1-\hat{p}_i^0)^{(1-f_i)}\,(1-\hat{p})}\right) \geq 0$$

$$= \log\left(\prod_{i=1}^{d} \left(\frac{\hat{p}_i^1}{\hat{p}_i^0}\right)^{f_i} \left(\frac{1-\hat{p}_i^1}{1-\hat{p}_i^0}\right)^{(1-f_i)} \cdot \frac{\hat{p}}{1-\hat{p}}\right) \geq 0$$

$$= \sum_{i=1}^{d} \left(f_i \underbrace{\log\left(\frac{\hat{p}_i^1}{\hat{p}_i^0}\right)} + (1-f_i) \underbrace{\log\left(\frac{1-\hat{p}_i^1}{1-\hat{p}_i^0}\right)} + \log\left(\frac{\hat{p}}{1-\hat{p}}\right)\right) \geq 0$$

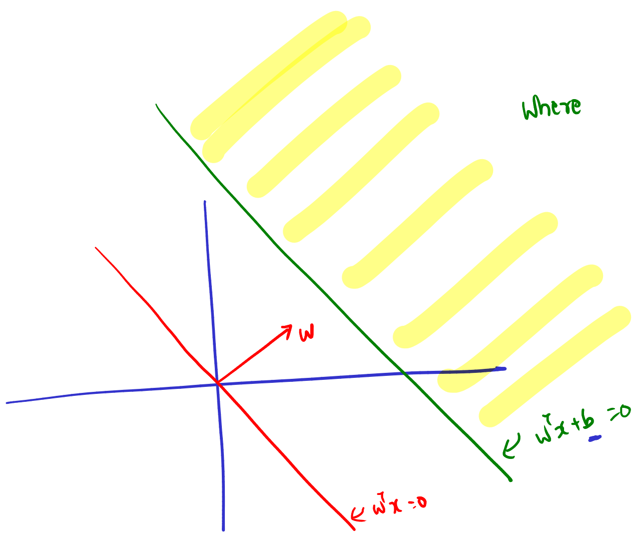
$$= \sum_{i=1}^{d} \underbrace{f_i \left(\log\left(\frac{\hat{p}_i^1(1-\hat{p}_i^0)}{\hat{p}_i^0(1-\hat{p}_i^1)}\right)\right)} + \underbrace{\log\left(\frac{1-\hat{p}_i^1}{1-\hat{p}_i^0}\right) + \log\left(\frac{\hat{p}}{1-\hat{p}}\right)} \geq 0$$

$x_{test} = [f_1 \cdots f_d]$

DECISION FUNCTION is of the form

Predict $y_{test} = 1$ if $w^T x_{test} + b \geq 0$, $w \in \mathbb{R}^d$

where $w_i = \underbrace{\log\left(\frac{\hat{p}_i^1(1-\hat{p}_i^0)}{\hat{p}_i^0(1-\hat{p}_i^1)}\right)}$, $b =$



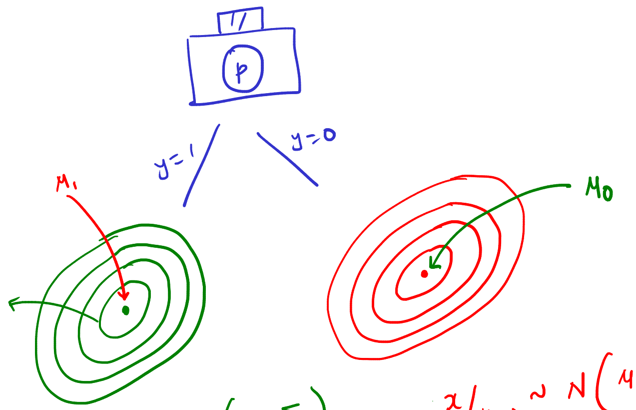
CONCLUSION:

→ DECISION FUNCTION OF NAIVE BAYES IS LINEAR!

DATA : $\{ (x_1, y_1), \dots \dots (x_n, y_n) \}$

$x_i \in \mathbb{R}^d \qquad y_i \in \{0, 1\}$

A GENERATIVE STORY



$x /_{y=1} \sim N(\mu_1, \Sigma) \qquad x /_{y=0} \sim N(\mu_0, \Sigma)$

PARAMETERS

- p
- μ_0, μ_1
- Σ

NOTE : In this model, covariances are assumed to be same

MAXIMUM LIKELIHOOD ESTIMATES

$$\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$$

← FRACTION OF points labelled 1.

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = 1) \cdot x_i}{\sum_{i=1}^{n} \mathbb{1}(y_i = 1)}$$

← Sample mean of data points labelled 1.

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} \mathbb{1}(y_i = 0) \cdot x_i}{\sum_{i=1}^{n} \mathbb{1}(y_i = 0)}$$

← Sample mean of data points labelled 1.

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}_{y_i}\right)\left(x_i - \hat{\mu}_{y_i}\right)^T$$

PREDICTION? Bayes rule.

$$P(y_{test} \mid x_{test}) \propto \underbrace{P(x_{test} \mid y_{test})}_{f(x_{test};\, \hat{\mu}_{y_{test}},\, \hat{\Sigma})} \cdot \underbrace{P(y_{test})}_{\hat{p}}$$

Predict $y_{test} = 1$ if

$$f(x_{test};\, \hat{\mu}_1, \hat{\Sigma}) \cdot \hat{p} \geq f(x_{test};\, \hat{\mu}_0, \hat{\Sigma}) \cdot (1-\hat{p})$$

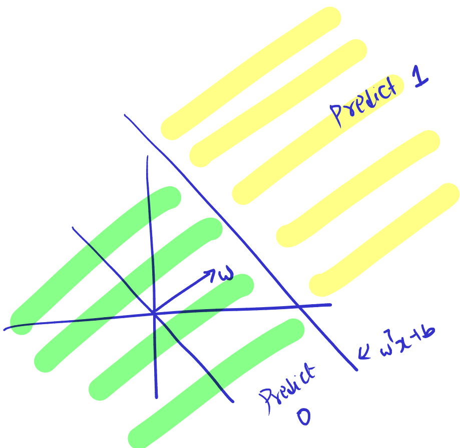
$$e^{-(x_{test}-\hat{\mu}_1)^T \hat{\Sigma}^{-1} (x_{test}-\hat{\mu}_1)} \cdot \hat{p} \geq e^{-(x_{test}-\hat{\mu}_0)^T \hat{\Sigma}^{-1} (x_{test}-\hat{\mu}_0)} (1-\hat{p})$$

On simplification [Take log]

Predict 1 if

$$\left(\left(\hat{\mu}_1 - \hat{\mu}_0\right)^T \hat{\Sigma}^{-1}\right) x_{test} + \underbrace{\hat{\mu}_0^T \hat{\Sigma}^{-1} \hat{\mu}_0 - \hat{\mu}_1^T \hat{\Sigma}^{-1} \hat{\mu}_1 + \log\left(\frac{1-\hat{p}}{\hat{p}}\right)} \geq 0$$

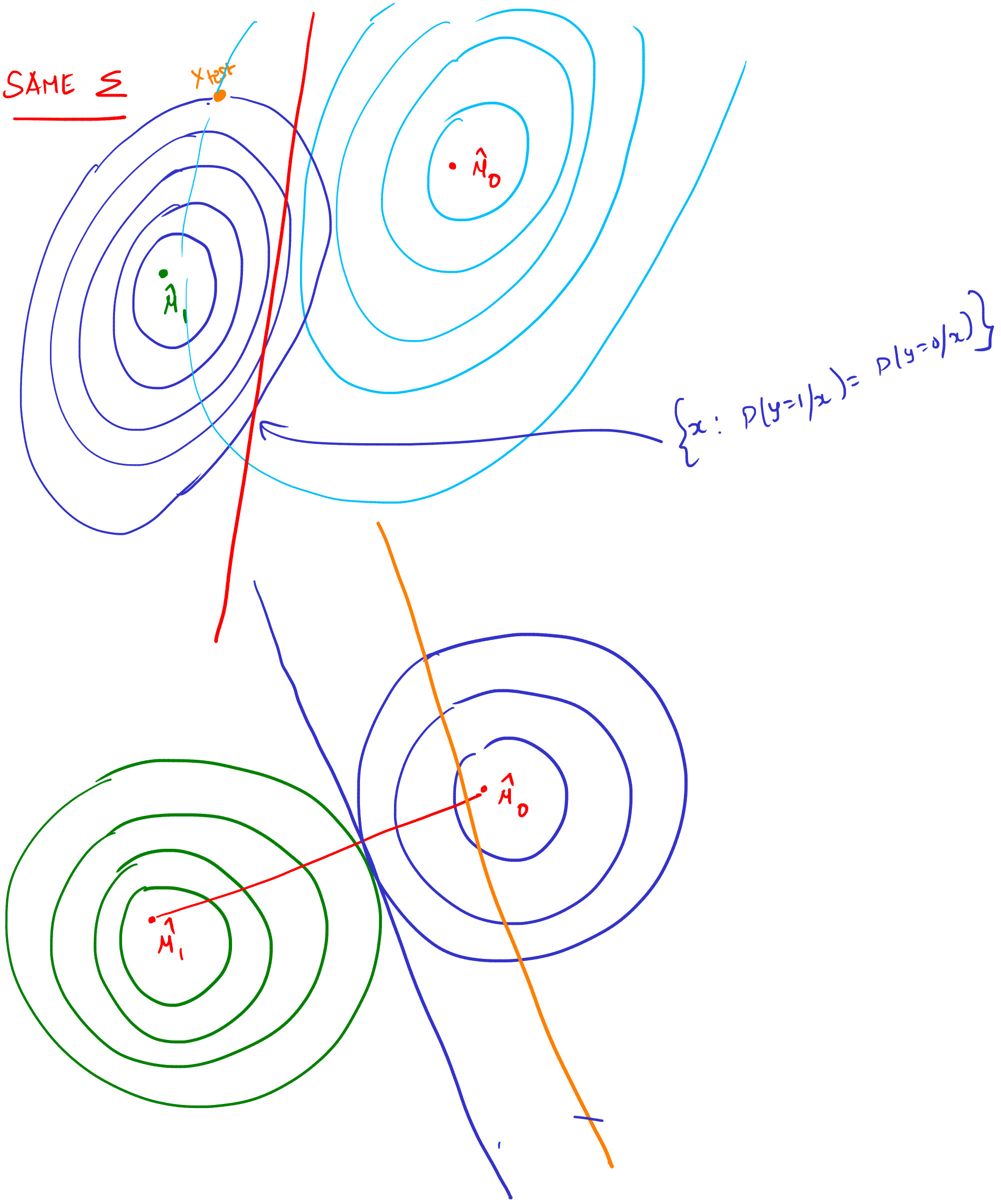
$$w^T x_{test} + b \geq 0$$



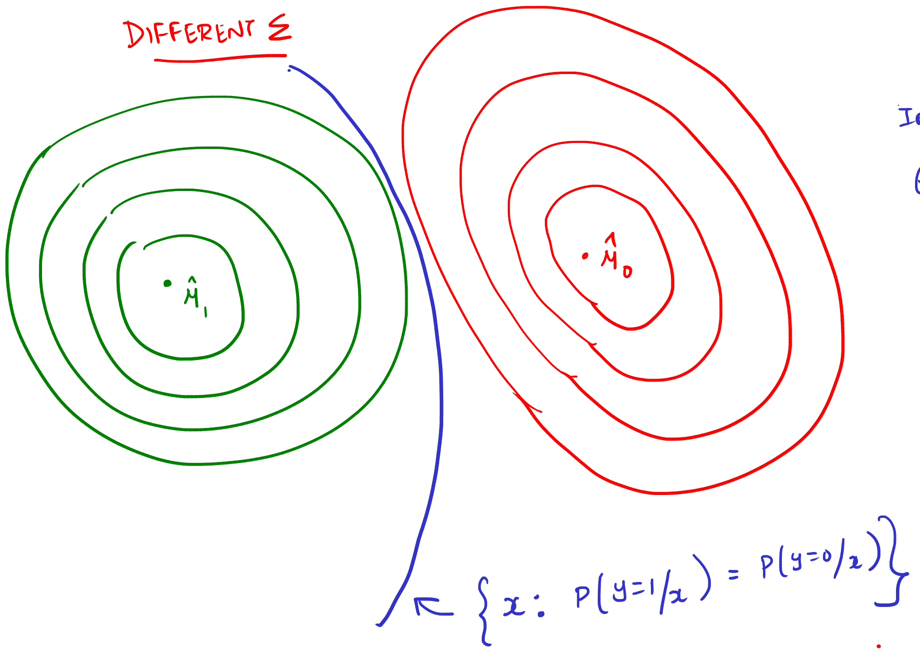
DECISION FUNCTION is LINEAR!

↳ Σ is same for both classes.

SAME Σ



DIFFERENT Σ



In general, Quadratic decision boundary.

GAUSSIAN NAIVE BAYES