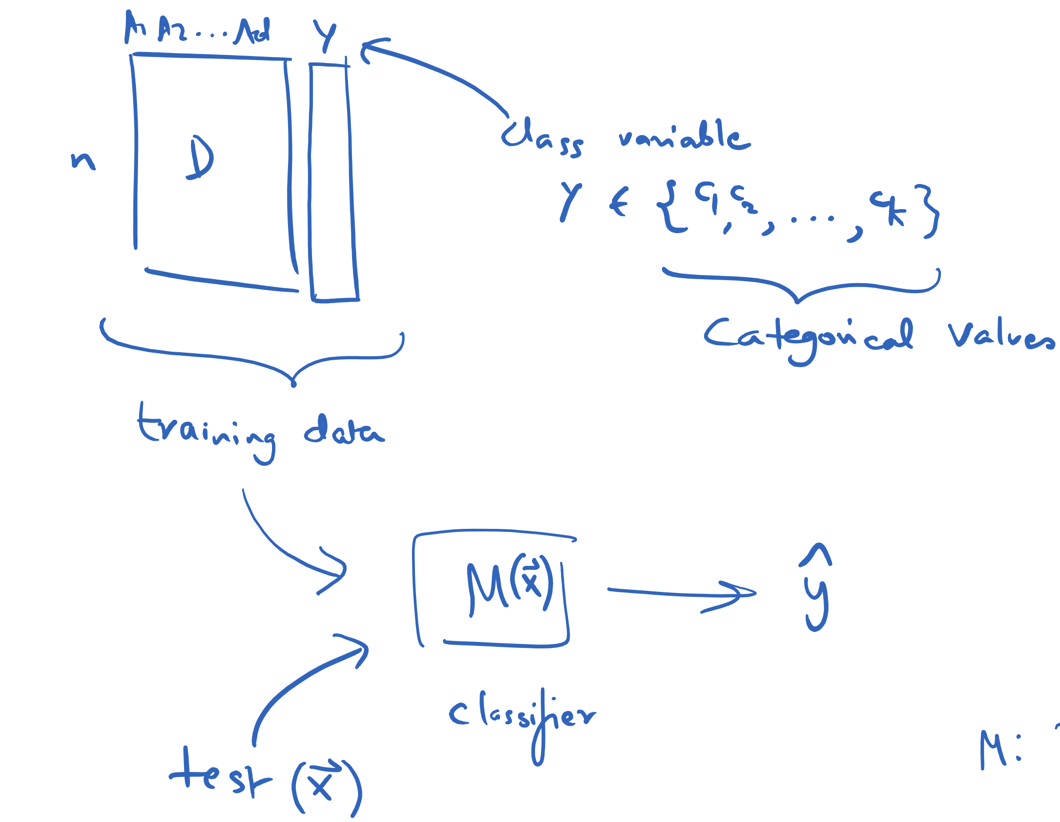
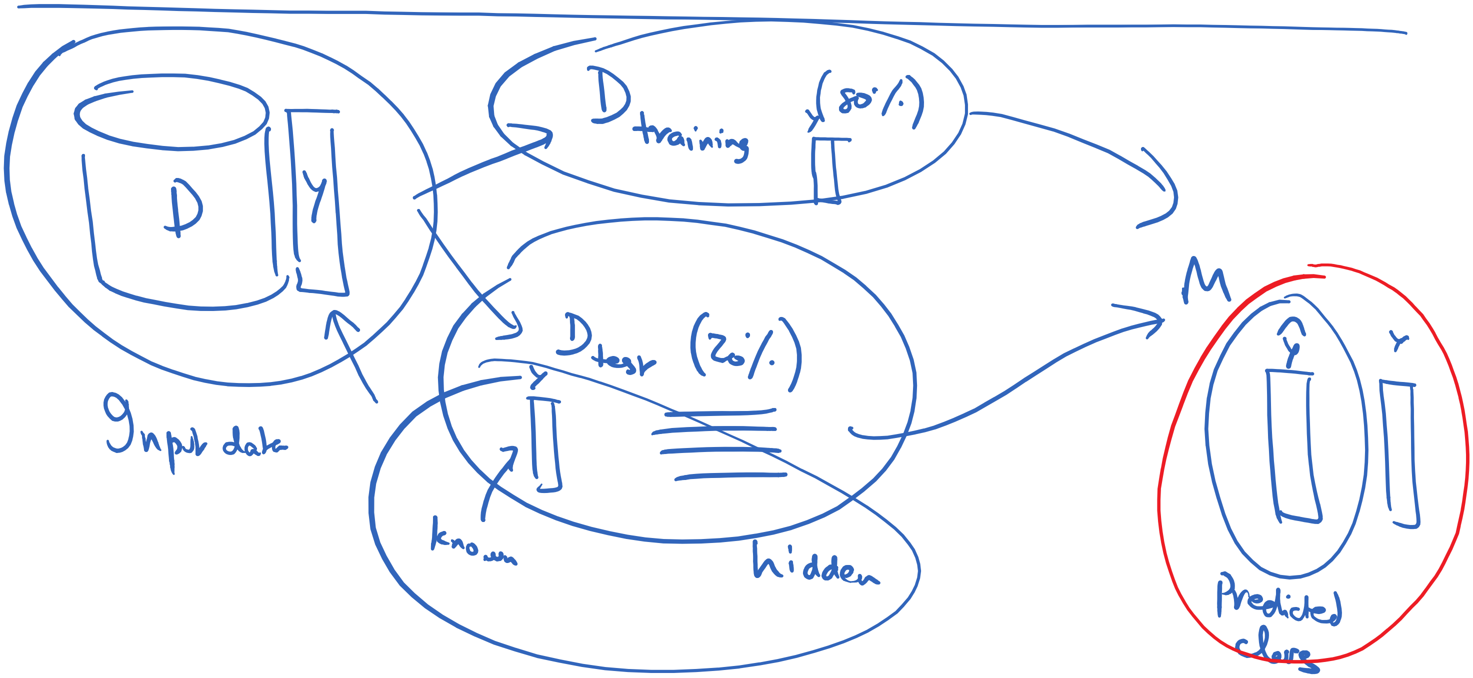
# Classification

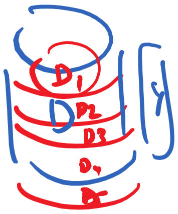



$M: \mathbb{R}^d \to \{c_1, c_2, \ldots c_k\}$




Cross-validation 20% test ≡ 5-fold cross validation

$D_1 \leftarrow$ test | $D_2 \leftarrow$ test | $D_3 \leftarrow$ test




$D_1 \leftarrow$ test | $D_2 \leftarrow$ test | $D_3 \leftarrow$ test

$\underbrace{D_2 : D_5 \text{ training}}_{\text{fold 1}}$ | $\underbrace{D_1, D_3 .. D_5}_{\text{fold 2}}$

$M_1$ | $M_2$

---

$D_{test} \rightarrow M(\vec{x}) \rightarrow \hat{y}$ vs $y$

$\hat{y}$: prediction

$y$: true class

0-1 loss

$$\frac{\#(\hat{y} \neq y)}{|D_{test}|} = \text{error rate}$$

per-class ?

credit card fraud

network intrusion detection

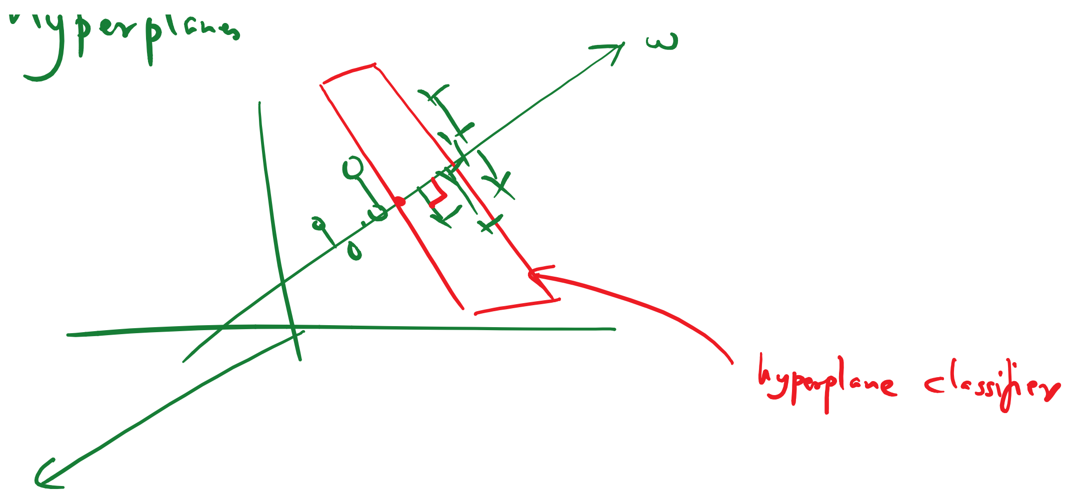
$\underbrace{\text{Pos}}$ fraud intrusion 0.1%

$\underbrace{\text{Neg}}$ Normal state 99.9%

Cost

---

hyperplanes




## Bayes classifier (Probabilistic)

training set D

$(\vec{x}_1, y_1)$
$(\vec{x}_2, y_2)$
$\vdots$
$(\vec{x}_n, y_n)$

$\vec{x} \in \mathbb{R}^d$

$y \in \underbrace{\{c_1, c_2, \dots, c_k\}}_{k \text{ classes}}$

given a "new" point $\vec{x}$, predict the class

Compute $P(c_i \mid \vec{x})$ ← ~~If known?~~

Bayes classifier → $h(\vec{x}) = \arg\max_{c_i} \{P(c_i \mid \vec{x})\}$

optimal classifier

$\underbrace{P(c_i \mid \vec{x})}_{\text{Posterior (probability}}$ vs $P(c_i)$

prior probability → before

probability

after

before

Bayes Theorem

$$\underbrace{P(C_i \mid \vec{x})} = \frac{P(C_i \,\&\, \vec{x})}{P(\vec{x})}$$

$$P(\vec{x} \mid c_i) = \frac{P(\vec{x} \,\&\, c_i)}{P(c_i)}$$

$$\boxed{P(c_i \mid \vec{x}) = \frac{P(\vec{x} \mid c_i) \cdot P(c_i)}{P(\vec{x})}}$$

Posterior

likelihood

Prior

$$P(c_i \mid \vec{x}) = \frac{P(\vec{x} \mid c_i) \cdot P(c_i)}{\sum_{a=1}^{k} P(\vec{x} \mid c_a) \cdot P(c_a)} \qquad \sum_{a=1}^{k} P(\vec{x} \mid c_a) \cdot P(c_a) = P(\vec{x})$$

$$\hat{P}(c_i) = \frac{n_i}{n} = \frac{\text{\# of points with class } i}{\text{size of training set}}$$

$c_1$, $c_2$, $c_1$, $c_2$, D

training sample

$\underbrace{P(\vec{x} \mid c_i)}_{\text{likelihood?}}$

→ parametric approach

1) Multivariate Normal Distribution ← GMMs

$\rightarrow \underbrace{\vec{\mu}_i, \Sigma_i}_{k \text{ pairs}}$

$\underbrace{\text{GMMs}}$ Gaussian Mixture Models

→ non-parametric

$\vec{x} \in \mathbb{R}^d$

$$D_1 = \{ \vec{x}_i \mid y_i = c_1 \} \rightarrow \hat{\mu}_1, \hat{\Sigma}_1$$


$D_1 = \{ x_i \mid y_i = c_1 \} \rightarrow \mu_1, \Sigma_1$

$D_2 = \quad \vdots \qquad\qquad \vdots$

$\vdots$

$D_k = \{ \vec{x}_i \mid y_i = c_k \} \rightarrow \hat{\mu}_k, \hat{\Sigma}_k$

d

$d = 1000$

we cannot do reliable estimate

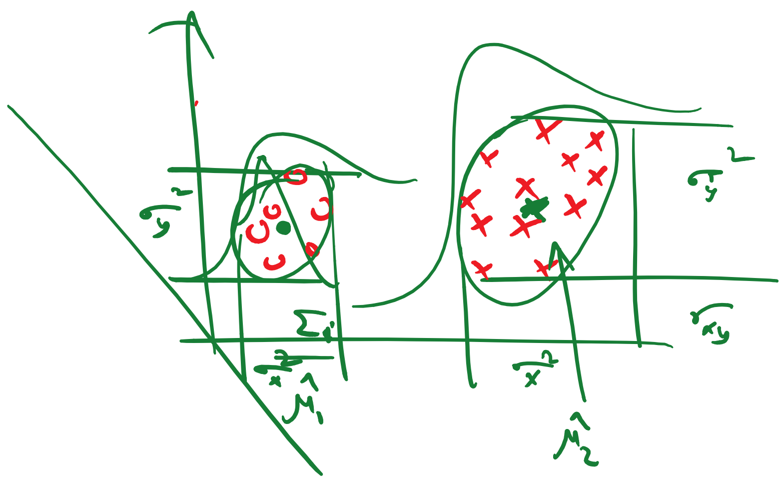
$\hat{\mu}_i$ — $d$ parameters $\times k$

$\hat{\Sigma}_i$ — $d \times d$ parameters $\times k$

$P(\vec{x} \mid c_i)$

$$P(c_i \mid \vec{x}) = \frac{N(\vec{x} \mid \underline{\hat{\mu}_i}, \underline{\Sigma_i}) \cdot P(c_i)}{P(x)}$$

$$\hat{y} = \underset{N}{\arg\max}\; P(c_i \mid \vec{x}) = \underset{c_i}{\arg\max} \left\{ N(\vec{x} \mid \hat{\mu}_i, \Sigma_i) \cdot P(c_i) \right\}$$




$$\Sigma_i = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

---

"Naive" Bayes : all attributes are independent

$$P(\vec{x} \mid c_i) = \prod_{j=1}^{d} \underbrace{P(x_j \mid c_i)}_{\text{Product of } d \text{ univariate normals}}$$

$$\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$$

$$x = \begin{pmatrix} x_2 \\ \vdots \\ x_d \end{pmatrix} \qquad j=1$$

Product of $d$ univariate normals

all Covariances are assumed to be 0

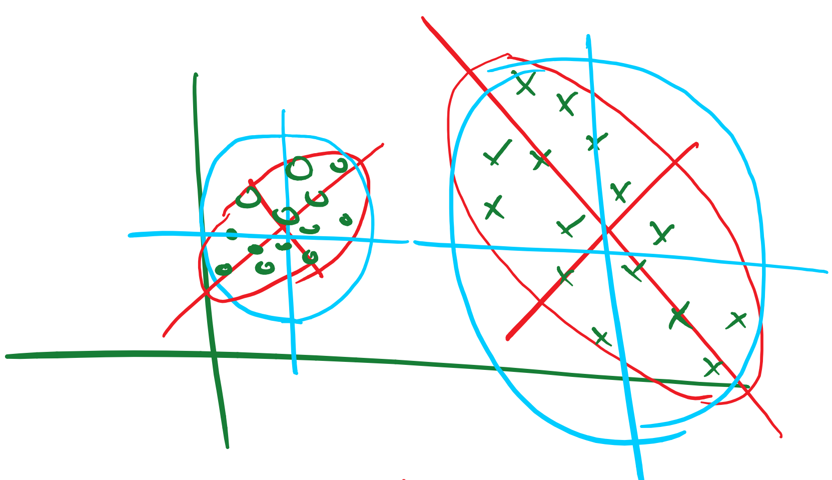
Full Bayes

$$\Sigma_i = \underbrace{\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & & & \end{pmatrix}}_{\frac{d(d-1)}{2}\ \text{distinct parameters}}$$

Naive Bayes

$$\Sigma_i = \underbrace{\begin{pmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 \ddots & \\ 0 & & \sigma_d^2 \end{pmatrix}}_{d\ \text{variances}}$$

$$\underbrace{P(\vec{x} \mid c_i)}_{\text{Joint}} = \underbrace{\prod_{j=1}^{d} P(x_j \mid c_i)}_{\text{independence}} = \prod_{j=1}^{d} N(x_j \mid \mu_i, \sigma_i^2)$$



Full Bayes Modeling

Naive Bayes

full Bayes Modeling

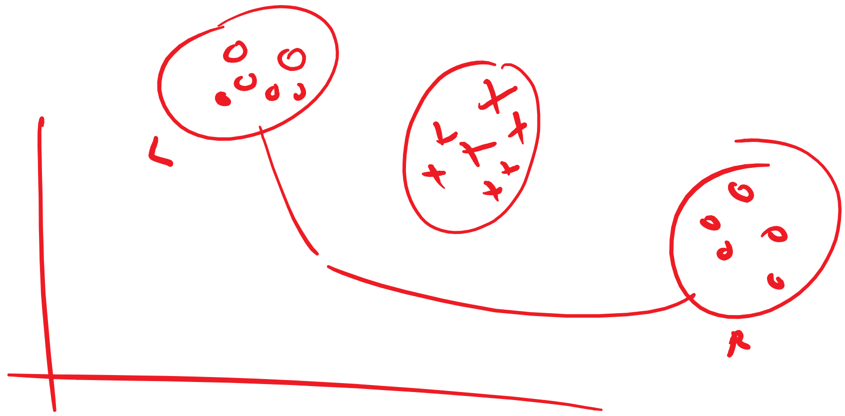
$\Downarrow$

full ellipsoids with rotation

Naive Bayes

$\Downarrow$

axis-aligned ellipsoids




$$P(x|c_0) : \alpha_L \cdot \underline{N(x|\mu_L, \Sigma_L)} + \alpha_R \cdot \underline{N(x|\mu_R, \Sigma_R)}$$

$$\alpha_R = 1 - \alpha_L$$

## Categorical Attributes


all attributes are categorical.

$P(\vec{x}|c_i) \leftarrow$ joint prob in d-dim

$$P(\vec{x}|c_i) = \prod_{j=1}^{d} P(x_j|c_i) \leftarrow \text{Independent}$$

## One Categorical Attribute

| | A |
|---|---|
| $x_1$ | H |
| $x_2$ | H |
| ⋮ | L |
| | L |
| ⋮ | H |
| | L |
| $x_n$ | ⋮ |

$A \in \{H, L\}$

If $A = H$, $v = 1$

$A = L$, $v = 0$

$x_n$ H L :

$A = L, \quad \underbrace{v = 0}$

Bernoulli Variable

$p_0, \quad p_1$

$1 - p_1$     unknown

Binomial Distribution (Counts)

$$P(\underset{\text{ones}}{K}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Expected # of 1's $= n \cdot p$

Sample

$$\boxed{\hat{p} = \frac{n_H}{n}}$$

expected # of H's $\quad n \cdot \frac{n_H}{n} = n_H \longleftarrow$

Variance : $n \cdot p \cdot (1-p)$

$\sigma^2$ : $\boxed{n \cdot \hat{p}(1-\hat{p})}$

---

$A(\text{risk}) \in \{H, M, L\}$

| A | $A_H$ $A_M$ $A_L$ |
|---|---|
| H | 1 0 0 |
| M | 0 1 0 |
| H | 1 0 0 |
| L | 0 0 1 |
| : | |
| : | |

$M = 3$

One-hot encoding

| A | | $A_H$ $A_M$ $A_L$ |
|---|---|---|
| H | $\rightarrow$ | 1 0 0 |
| M | $\rightarrow$ | 0 1 0 |
| L | $\rightarrow$ | 0 0 1 |

$p_H \quad p_M \quad p_L$

Multivariate Bernoulli

$\hat{p} \quad n_H \quad n_M$

m outcome

$$\hat{p}_H = \frac{n_H}{n} \qquad \underset{\hat{p}_m}{\frac{n_m}{n}} \qquad \frac{n_L}{n} = \hat{p}_L$$

Bernoulli

$$p_H + p_m + p_L = 1$$

Multinomial Distribution

$$\boxed{P(n_H, n_m, n_L) = \binom{n}{n_H, n_m, n_L} \cdot p_H^{n_H} \cdot p_m^{n_m} \cdot p_L^{n_L}}$$

$$= \left(\frac{n!}{n_H!\, n_m!\, n_L!}\right) \cdot \text{''}$$

| | $A_1$ | $A_2$ |
|---|---|---|
| $x_1$ | H | A |
| $x_2$ | L | B |
| | H | C |
| | L | A |
| | ⋮ | B |
| | | C |

$|A_1| = m_1 \qquad |A_2| = m_2$

$m_1 = 2 \qquad m_2 = 3$

$A_1 \in \{H, L\}$

$A_2 \in \{A, B, C\}$

$A_1 \to A_H : 10$, $A_L : 01$

$A_2 \to A_A : 100$, $A_B : 010$, $A_C : 001$

$$\vec{x}_1 = (H, A) = \begin{array}{ccccc} A_1' & A_2' & A_3' & A_4' & A_5' \end{array}$$

| | |
|---|---|
| 1 0 | 1 0 0 |
| 0 1 | 0 1 0 |
| ⋮ | ⋮ |

$\hat{p}_H : n_H/n$

$\hat{p}_L : n_L/n$

$A_1$

$\hat{p}_A : n_A/n$

$\hat{p}_B : n_B/n$

$\hat{p}_C : n_C/n$

$A_2$

joint probability

$A_1$ & $A_2$

| $A_1$ \ $A_2$ | A | B | C | |
|---|---|---|---|---|
| H | $n_{HA}$ | $n_{HB}$ | $n_{HC}$ | $n_H$ |
| L | $n_{LA}$ | $n_{LB}$ | $n_{LC}$ | $n_L$ |

| | | | | |
|---|---|---|---|---|
| L | $n_{LA}$ | $n_{LB}$ | $n_{LC}$ | $n_L$ |
| | $n_A$ | $n_B$ | $n_C$ | $n$ |

Contingency table

$A_1$

$A_2$

$$\hat{P}_{HA} = \frac{n_{HA}}{n}$$

$\underbrace{\text{joint probability estimation}}_{\text{unreliable}}$

---

$P(x \mid c_1)$

$x \in \{0,1\}^d$

Categorical

full Bayes

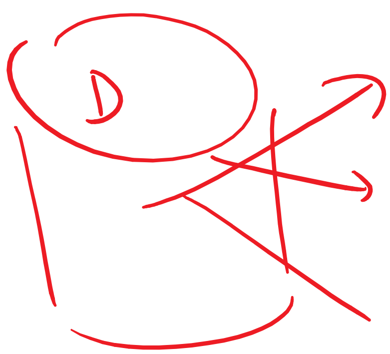
$x = (H, M, A)$

| | | |
|---|---|---|
| H | H | A |
| L | M | B |
| | L | C |
| | | D |

$2 \times 3 \times 4$

24

$$\hat{P}_{HMA} = n_{HMA} / n$$



$D_1 \longrightarrow \hat{P}_{HMA \mid c_1}$

$D_2 \quad \hat{P}_{HMA \mid c_2}$

$\vdots$

$D_k$

naive Bayes

$$P(H, M, A \mid c_1) = P(H \mid c_1) \cdot P(M \mid c_1) \cdot P(A \mid c_1)$$