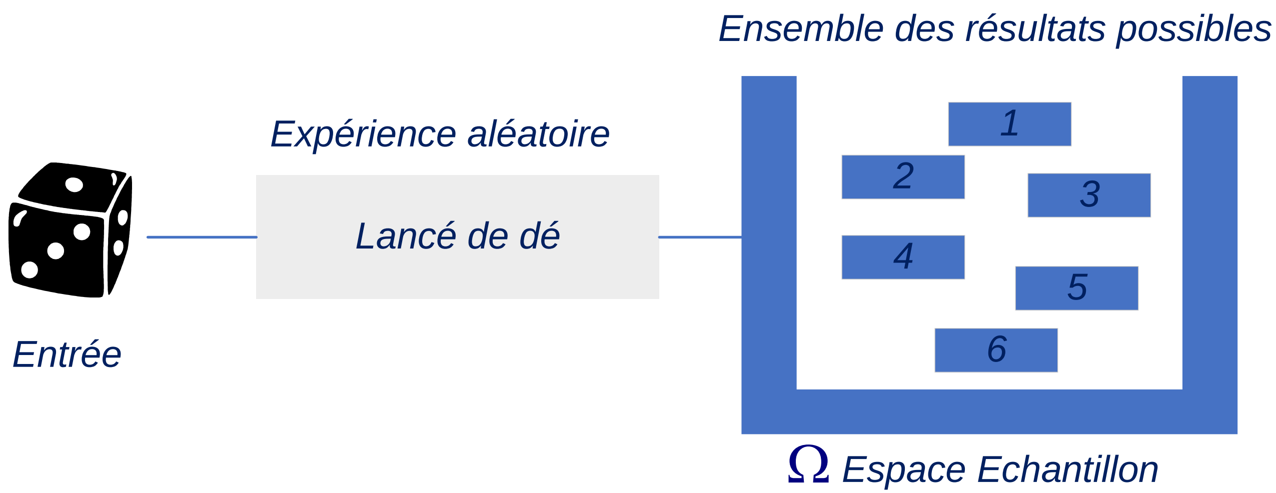
ANALYSE DISCRIMINANTE

CLASSIFICATION BAYESIENNE



- Expérience aléatoire
  Action ou un processus qui engendre une observation dont on ne peut pas prédire à l'avance le résultat

- Espace échantillon
  L'ensemble des résultats possibles issu d'une expérience aléatoire est appelé espace échantillon
  - Un espace échantillon est discret si l'ensemble des éléments qu'il contient est fini ou infini dénombrable
  - Un espace échantillon est continu si l'ensemble des éléments qu'il contient est infini non dénombrable

- Evènement
  Proposition logique concernant le résultat d'une expérience
  On peut donc considérer qu'un évènement est un sous ensemble de l'espace échantillon (issu d'une expérience aléatoire)

Assertion effectuée a priori : l'expérimentateur décide de s'intéresser à une problématique définie

- Si l'on considère simultanément la réalisation de deux évènements E1 et E2, il est possible d'effectuer des opérations sur ces ensembles

Exemple : Lors d'un lancé d'un dé, l'expérimentateur décide de s'intéresser aux évènements suivants:

- Obtenir un multiple de 3 : E1={3,6} → *Évènement E1*
- Obtenir un nombre pair E2={2,4,6} → *Évènement E2*

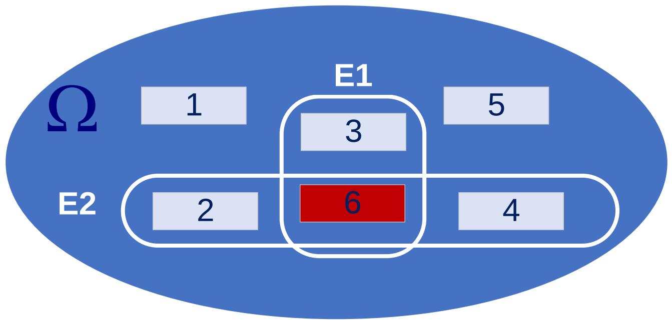
Évènements : propositions logique concernant le résultat d'une expérience

- Pour rappel
  - Évènement simple: ne contient qu'un seul résultat
  - Évènement certain : contient tous les résultats
  - Évènement impossible : ensemble vide

<u>Les évènements sont des sous ensembles de</u> Ω

⇒ Intersection

On appelle intersection de deux évènements E1 et E2, l'événement qui est réalisé si et seulement si E1 et E2 le sont. Il est donc constitué des **éventualités** appartenant à la fois à E1 et E2.

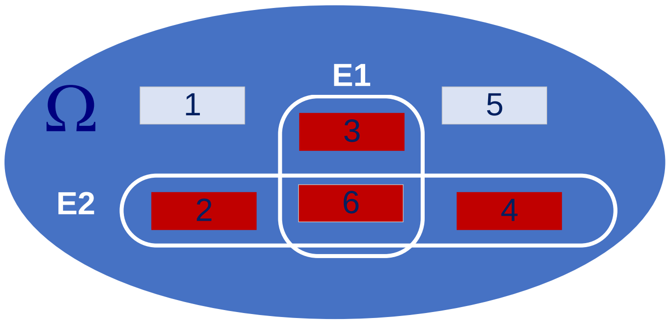


- Propriétés de l'intersection

  - Deux évènements E1 , E2 sont disjoints s'ils ne peuvent être réalisés simultanément. On a alors $E_1 \cap E_2 = \varnothing$
  - $E_1 \cap \overline{E}_1 = \varnothing \longrightarrow$ Évènements incompatibes
  - $\Omega \cap E_1 = E_1 \longrightarrow$ Élément neutre
  - $\varnothing \cap E_1 = \varnothing \longrightarrow$ Élément absorbant
  - $E_1 \cap E_2 = E_2 \cap E_1 \longrightarrow$ Commutativité
  - $(E_1 \cap E_2) \cap E_3 = E_1 \cap (E_2 \cap E_3) \longrightarrow$ Associativité
  - $(E_1 \cap E_2) \cap E_3 = (E_1 \cap E_3) \cup (E_2 \cap E_3) \longrightarrow$ Transitivité

$\Rightarrow$ **Réunion**

On appelle réunion de deux évènements E1 et E2, l'évènement qui est réalisé si et seulement si E1 ou E2 est réalisé

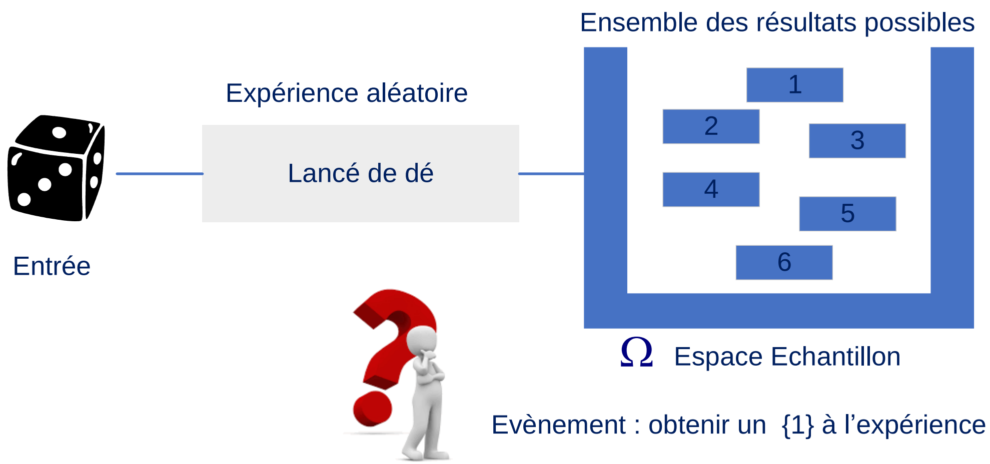


- Propriétés de la réunion
  - $E_1 \cup \overline{E}_1 = \Omega$
  - $\Omega \cup E_1 = \Omega$
  - $\varnothing \cup E_1 = E_1$
  - $E_1 \cup E_2 = E_2 \cup E_1$
  - $(E_1 \cup E_2) \cup E_3 = E_1 \cup (E_2 \cup E_3)$
  - $(E_1 \cup E_2) \cup E_3 = (E_1 \cup E_3) \cap (E_2 \cup E_3)$

Probabilité (simple)

⇒ Définitions

- La probabilité est utilisée pour quantifier la vraisemblance d'un évènement issu d'une expérience aléatoire



Cas favorables obtention d'un 1

Réalisation de l'expèrience

Cas défavorableS: {1,2,3,4,5}

Nombre de cas favorables ne = 1

Nombre de cas possibles (ensemble des résultats possibles) N = 6

- Définition classique de la probabilité (dite aussi probabilité a priori)
  - La probabilité d'un évènement E est le rapport entre le nombre de cas favorables (n_e) et le nombre de cas possibles (N) **tous également vraisemblables**

Expérience aléatoire

probabilité $\longrightarrow$ $P(E=1)=\dfrac{n_e}{N}=\dfrac{1}{6}$ $\longrightarrow$ résultat de l'expérience

évènement

- Approche fréquentiste
  - Elle est fondée sur la répétition (N) de l'expérience aléatoire. A chaque essai, on note le résultat de l'épreuve. Soit ne le nombre d'apparitions de l'évènement E (ex : résultat = 1)
  - L'approche fréquentiste est utilisée lorsque le décompte du nombre de cas favorables ou le nombre de cas possibles est complexe ou impossible. Il s'agit d'une approche expérimentale

On effectue n = 1000 lancés de dé

succés	*Lancé du dé*
1	*5*
1	*9*
1	*22*
1	*23*
1	*30*
1	*32*
1	*33*
Obtention d'un '1' au 5 ième lancé	
1	*9984*
1	*9985*
1	*9987*
1	*9988*
1	*9990*
1	*9994*

n_S	*N*
1	*5*
2	*9*
3	*22*
4	*23*
5	*30*
6	*32*
7	*33*
1686	*9966*
1687	*9972*
1688	*9984*
1689	*9985*
1690	*9987*
1691	*9988*
1692	*9990*
1693	*9994*
somme	

La probabilité que l'évènement se réalise est :

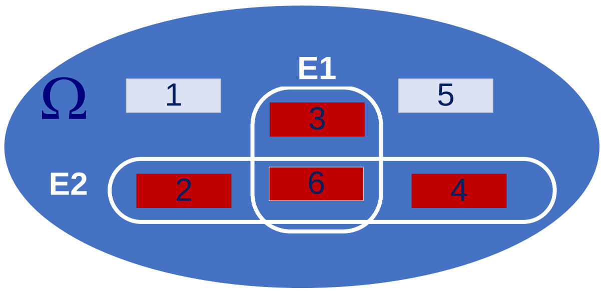
$$\lim_{N\to\infty}\left(\frac{n_e}{N}\right)=P(E)$$

$$\mathrm{P}(E)=\lim_{n\to\infty}F_n(E)=\frac{n_s}{n}$$

- La probabilité de survenue d'un évènement E est un nombre compris entre 0 et 1: $0 \leq P(E) \leq 1$
- La probabilité d'un évènement certain = 1: $P(\Omega) = 1$ $P(\varnothing) = 0$
- Soit évènement complémentaire: $P(\bar{E}) = 1 - P(E)$
- Si E1 et E2 sont deux évènements **incompatibles** $E_1 \cap E_2 = \varnothing$ alors la probabilité de réalisation de l'un **ou** de l'autre est égale à la somme des probabilités $P(E_1 \cup E_2) = P(E_1) + P(E_2)$
- Si E1 et E2 sont deux évènements **quelquonques** alors la probabilité de réalisation de l'un **ou** de l'autre est égale à : $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
- Si E1 et E2 sont deux évènements **quelquonques** alors la probabilité de réalisation de l'un **ou** de l'autre est égale à : $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$
- Si E1 et E2 sont deux évènements indépendants alors: $P(E_1 \cap E_2) = P(E_1) \times P(E_2)$

Attention ! Il ne faut pas confondre indépendance statistique (ex : tirage avec remise) et événements incompatibles

Exemple Obtenir un nombre pair ou un multiple de 3 $\longrightarrow$ $P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$



Obtenir un nombre pair E2={2,4,6}

Obtenir un multiple de 3 : E1={3,6}

$$P(E_1) = \frac{2}{6} \qquad P(E_2) = \frac{3}{6} \qquad P(E_1 \cap E_2) = P(E_1) \times P(E_2) = \frac{3}{6} \times \frac{2}{6} = \frac{1}{6}$$

$$P(E_1 \cup E_2) = \frac{2}{6} + \frac{3}{6} - \frac{1}{6} = \frac{4}{6}$$

- Soit deux évènements E1 et E2. On appelle probabilité conditionelle de E1 par rapport à E2, la probabilité de réalisation de l'évènement E1 **sachant que l'évènement E2 s'est réalisé**

E1 et E2 sont dépendants $$P(E_1 / E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

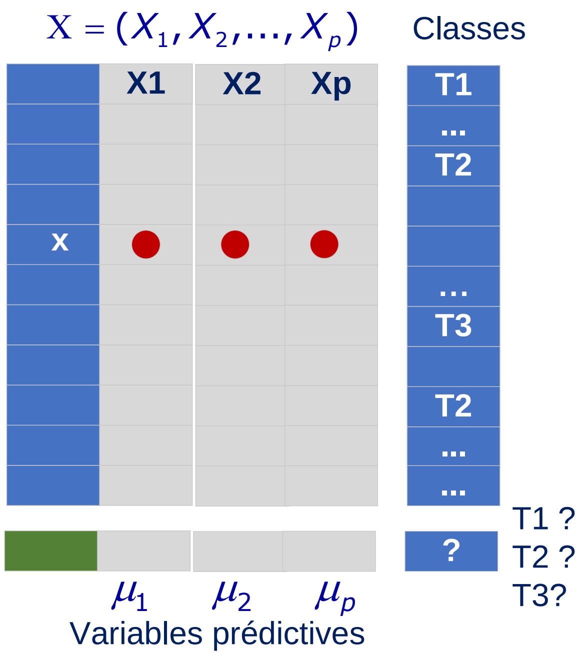
Dans le cas où E1 et E2 sont indépendants

$$P(E_1 / E_2) = P(E_1)$$
$$P(E_2 / E_1) = P(E_2)$$
$$P(E_1 \cap E_2) = P(E_1) \times P(E_2)$$

- Le théorème de Bayes est un outil de modélisation

$$\left.\begin{aligned} P(E_1 / E_2) &= \frac{P(E_1 \cap E_2)}{P(E_2)} \\ P(E_2 / E_1) &= \frac{P(E_2 \cap E_1)}{P(E_1)} \end{aligned}\right\} P(E_1 / E_2) = \frac{P(E_2 / E_1) \times P(E_1)}{P(E_2)}$$

Nous verrons comment utiliser ce théorème et sa signification lors du prochain chapitre///



Obtenir les meilleures séparations possibles entre les classes

↓

Classifier de nouvelles données dans les classes (prédéfinies)

Analyse multidimensionnelle

$\Rightarrow$ X est une matrice (composée de x colonnes, chaque colonne correspondant à une variable)

$\Rightarrow$ La moyenne $\mu = (\mu_1, \mu_2, \ldots, \mu_p)$ Dimension (p x 1)

$\Rightarrow$ La Matrice des variances-covariances Dimension (p x p)

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & \\ \sigma_{12}^2 & \sigma_2^2 & \\ \sigma_{13}^2 & \sigma_{23}^2 & \sigma_p^2 \end{pmatrix}$$

Hypothèse forte

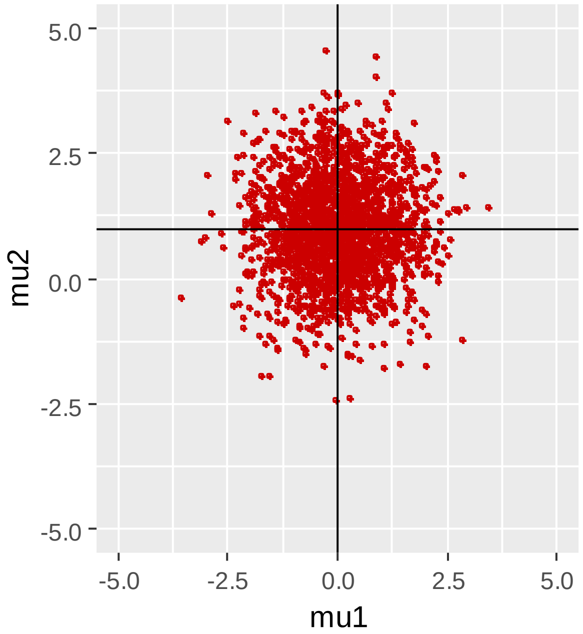
Multinormalité : La distribution de probabilité de X (multivariée) sous hypothèse de Normalité

$$\mathrm{X} \rightarrow N(\mu, \Sigma) \longrightarrow f_X(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{-1/2}} e^{\left(-1/2(x-\mu)^t \Sigma^{-1}(x-\mu)\right)}$$

Pour rappel, si p = 1 On retrouve la loi normale univariée

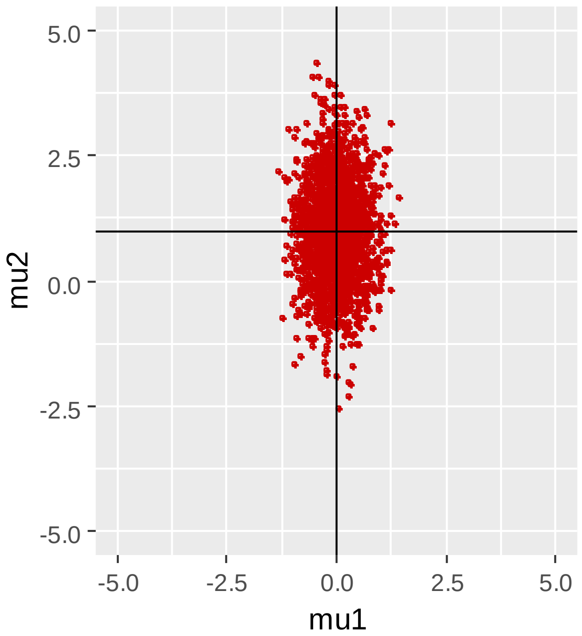
$$f_{p=1}(x) = \frac{1}{(2\pi\sigma_1^2)^{1/2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right)$$

Exemple de distributions multivariées pour p = 2 $\mu = (\mu_1, \mu_2)$



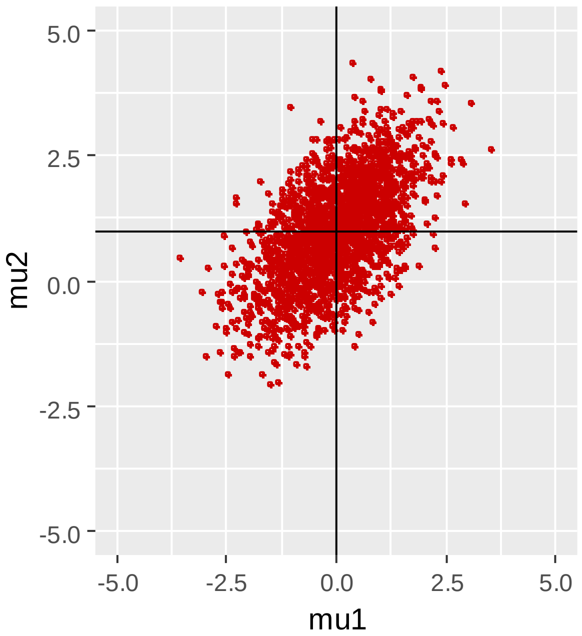
$\mu = (0,1)$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$\mu = (0,1)$

$$\Sigma = \begin{pmatrix} 0.2 & 0 \\ 0 & 1 \end{pmatrix}$$

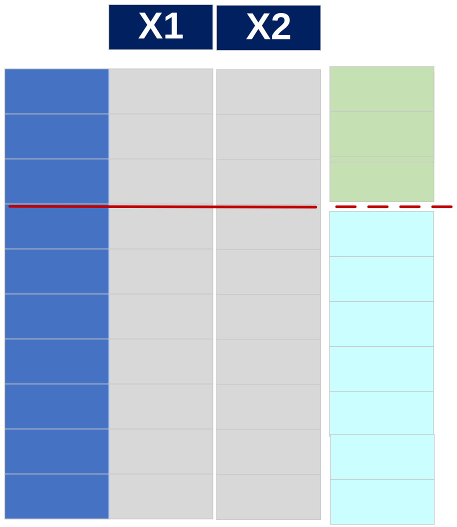


$\mu = (0,1)$

$$\Sigma = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}$$

- **Cas d'un mélange Gaussien (représentation)**

Exemple: p = 2 et deux classes



$X_1 = (X_{1,k=1}, X_{2,k=1})$

k =1

$X_1 \rightarrow N(\mu_{k=1}, \Sigma_{k=1})$

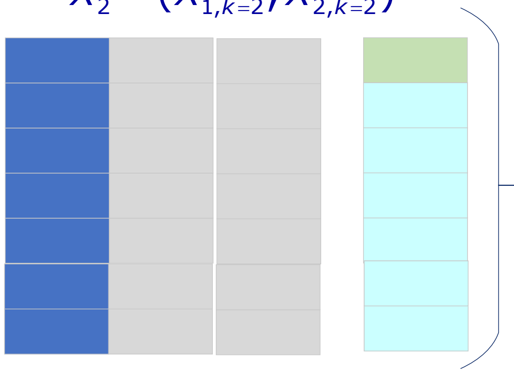
$\mu_{k=1} = (0,0)$

$$\Sigma_{k=1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Soit $\pi = 0.3$: proportion d'observations pour k = 1

Soit $1-\pi = 0.7$: proportion d'observations pour k = 2

$X_2 = (X_{1,k=2}, X_{2,k=2})$



k = 2

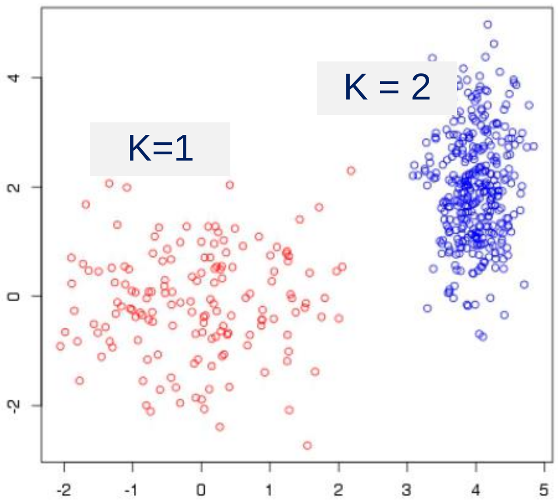
$X_2 \rightarrow N(\mu_{k=2}, \Sigma_{k=2})$

$\mu_{k=2} = (4,2)$

$$\Sigma_{k=2} = \begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}$$

La densité de probabilité du mélange est :

$$f(X) = \pi f(X_{k=1}) + (1-\pi) f(X_{k=2})$$



- Généralisation : la densité de probabilité d'un mélange Gaussien est :

$$f(X) = \sum_{k=1}^{k} \pi_k f(X_k) \quad \text{avec} \quad \sum_{k=1}^{K} \pi_k = 1$$

Sous la condition de multinormalité $X_k \to N(\mu_k, \Sigma_k)$

$$f(X_k) = \frac{1}{(2\pi)^{p/2} \left|\Sigma_K\right|^{-1/2}} e^{\left(-1/2 (x-\mu_K)^t \Sigma_K^{-1} (x-\mu_K)\right)}$$

$\mathrm{X} = (X_1, X_2, \ldots, X_p)$

	X1	X2	Xp
X	x_1	x_2	x_p

Classe C

k = 1
…
k = 2
k =
…
…
…

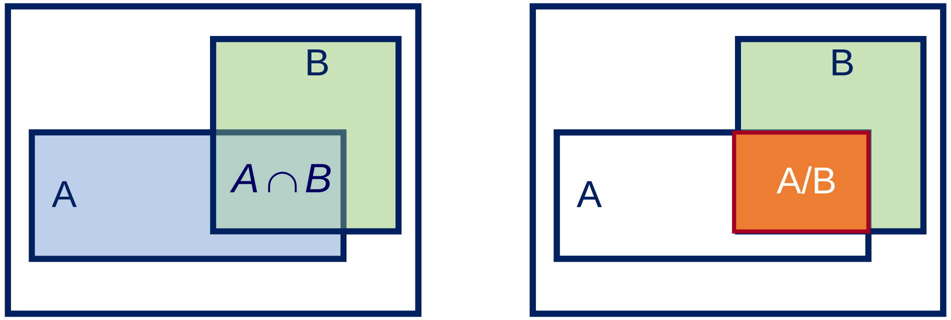
Correspond à la probabilité pour un individus d'observer les valeurs x_1, x_2, x_p sachant qu'il appartient à la classe k

$$P(X = x / Y = C_k) = f(X_k) = \frac{1}{(2\pi)^{p/2} \left|\Sigma_k\right|^{-1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1} (x-\mu_k)\right)$$

- Rappels: Probabilité conditionnelle et règle de Bayes

Soit deux évènements A et B. On appelle probabilité conditionelle de A par rapport à B, la probabilité de réalisation de l'évènement A sachant que l'évènement B s'est réalisé

P(A / B) = probabilité conditionnelle de A sachant que B s'est réalisé
$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$



- Remarque : Deux évènements sont indépendants si la connaissance de l'un ne modifie pas la connaissance de l'autre

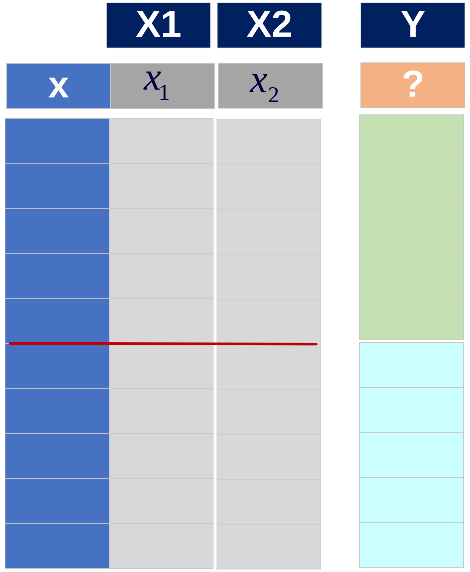
Si deux évènements A et B sont indépendants alors

$$P(A \cap B) = P(A) \times P(B) \quad \longrightarrow \quad \begin{array}{l} P(A/B) = P(A) \\ P(B/A) = P(B) \end{array}$$

Règle de Bayes

$$\left.\begin{array}{l} P(A/B) = \dfrac{P(A \cap B)}{P(B)} \\[2ex] P(B/A) = \dfrac{P(A \cap B)}{P(A)} \end{array}\right\} P(A/B) = \frac{P(B/A) \times P(A)}{P(B)}$$

- Utilisation de la règle de Bayes en analyse discriminante



Soit la variable de groupe $Y \in [C_1, \ldots C_k]$

Soit x, une observation que l'on cherche à classer

On Cherche donc à calculer la probabilité d'appartenance à une classe k connaissant les valeurs de x = (x_1, x_2)

$$P(Y = C_k / X = x)$$

D'après le théorème de Baye :

$$\underbrace{P(Y = C_k / X = x)}_{\text{Probabilité } a\ posteriori} = \frac{\underbrace{P(X = x / Y = C_k)}_{\text{Vraisemblance}} \times \underbrace{P(Y = C_k)}_{\text{Probabilité } a\ priori}}{P(X = x)}$$

Vraisemblance

$$P(X=x/Y=C_k)=f(X_k)=\frac{1}{(2\pi)^{p/2}\left|\Sigma_k\right|^{-1/2}}\exp\left(-\frac{1}{2}(x-\mu_k)^t\,\Sigma_k^{-1}(x-\mu_k)\right)$$

Probabilité *a priori*

$$P(Y=C_k)=\pi_k$$

$$P(Y=C_k\ /\ X=x)=\frac{P(X=x/Y=C_k)\times P(Y=C_k)}{P(X=x)}$$

$$P(X)=f(X)=\frac{1}{(2p)^{1/p}\left|\Sigma\right|^{-1/2}}\exp\left(\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right)$$

Distribution de probabilité multivariée indépendamment de la classe d'appartenance. La moyenne et la variance sont calculées indépendamment des classes

Pour un nouvel individu, chercher sa classe d'appartenance revient à **maximiser** $P(Y=C_k\ /\ X=x)$

- En pratique, on calculera pour cet individu:

$$\left.\begin{array}{l} P(Y=C_{k=1}\ /\ X=x) \\ P(Y=C_{k=2}\ /\ X=x) \\ \dots\dots\dots\dots\dots\dots \\ P(Y=C_{k=K}\ /\ X=x) \end{array}\right\} \quad \arg\max_k\left(P(Y=C_k\ /\ X=x)\right)$$

Remarque importante pour simplification : quelquesoit les classes P(X) est la même. Il s'agit donc d'une constante qui n'intervient pas lors de la maximisation

On distingue plusieurs cas

Cas 1: Homoscédasticité et uniformité

$\Sigma_K = \Sigma \; \forall_K$ Homoscédasticité : même variance pour toutes les classes

$\pi_k \approx \frac{1}{K}$ Uniformité: proportion d'observations identiques dans chaque classe

$$P(Y = C_k / X = x) = \frac{1}{k(2p)^{1/p} |\Sigma|^{-1/2}} \exp\left(\frac{1}{2} (x - \mu_k)^t \Sigma^{-1} (x - \mu_k) \right)$$

On cherche donc $\arg\max_k \left(P(Y = C_k / X = x) \right)$

Pour simplifier, on passe en log $\arg\max_k \left(\log\left(P(Y = C_k / X = x) \right) \right)$

$$\arg\max_k \left(\log\left(\frac{1}{k(2p)^{1/p} |\Sigma|^{-1/2}} \exp\left(\frac{1}{2} (x - \mu_k)^t \Sigma^{-1} (x - \mu_k) \right) \right) \right)$$

$$\arg\max_k \left(-\log\left(k(2p)^{1/p} \right) - \frac{1}{2} \log\left(|\Sigma| \right) - \frac{1}{2} \left((x - \mu_k)^t \Sigma^{-1} (x - \mu_k) \right) \right)$$

Les proportions et les variances étant identiques pour chaque classe elles ne dépendant pas de la classe

On cherche tout simplement :

$$\arg\max_k \left(-\frac{1}{2} \left((x - \mu_k)^t \Sigma^{-1} (x - \mu_k) \right) \right)$$

Cas 2: Homoscédasticité et non uniformité

$\Sigma_K = \Sigma \ \forall_K$ Homoscédasticité : même variance pour toutes les classes

La proportion d'observations est différente dans chaque classe

$$P(Y = C_k / X = x) = \frac{\pi_k}{(2p)^{1/p} |\Sigma|^{-1/2}} \exp\left(\frac{1}{2} (x - \mu_k)^t \Sigma^{-1} (x - \mu_k) \right)$$

$$\arg\max_k \left(\log\left(P(Y = C_k / X = x) \right) \right)$$

$$\arg\max_k \left(+ \log(\pi_k) - \log\left((2p)^{1/p} \right) - \frac{1}{2} \log\left(|\Sigma| \right) - \frac{1}{2} \left((x - \mu_k)^t \Sigma^{-1} (x - \mu_k) \right) \right)$$

$$\arg\max_k \left(-\frac{1}{2} \left((x - \mu_k)^t \Sigma^{-1} (x - \mu_k) \right) + \log(\pi_k) \right)$$

Score de Fisher

Cas 3: Hétéroscédasticité et non uniformité QDA

$$P(Y = C_k / X = x) = \frac{\pi_k}{(2p)^{1/p} |\Sigma_k|^{-1/2}} \exp\left(\frac{1}{2} (x - \mu_k)^t \Sigma_k^{-1} (x - \mu_k) \right)$$

$$\arg\max_k \left(\log\left(P(Y = C_k / X = x) \right) \right)$$

$$\arg\max_k \left(+ \log(\pi_k) - \frac{1}{2} \log\left(|\Sigma_k| \right) - \frac{1}{2} \left((x - \mu_k)^t \Sigma^{-1} (x - \mu_k) \right) \right)$$

Fonction discriminante quadratique

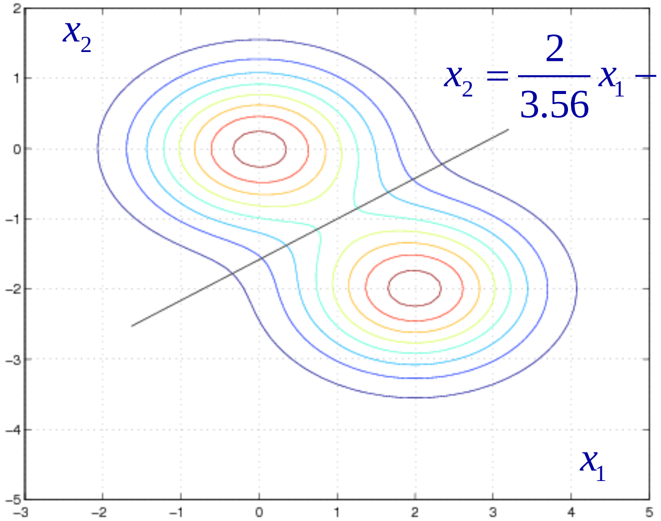
Exemple 1: Cas uniforme et même variance 2 groupes (k = 2) et 2 variables

$$\pi_1 = \pi_2 = 0.5$$

$$\mu_1 = (0,0)^t$$

$$\mu_2 = (2,-2)^t$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0,5626 \end{pmatrix}$$



On développe la fonction discriminante

$$\delta_k = -\frac{1}{2}\left((x-\mu_k)^t \Sigma^{-1}(x-\mu_k)\right) + \log(\pi_k) = x^t \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^t \Sigma^{-1} - \frac{1}{2} x^t \Sigma^{-1} x + \log(\pi_k)$$

La frontière entre les deux distributions (de la classe 1 et de la classe 2) est $\delta_{k=1} = \delta_{k=2}$

$$\log\left(\frac{\pi_{k=1}}{\pi_{k=2}}\right) - \frac{1}{2}\left(\mu_{k=1} - \mu_{k=2}\right)^t \Sigma^{-1}\left(\mu_{k=1} - \mu_{k=2}\right) + x^t \Sigma^{-1}\left(\mu_{k=1} - \mu_{k=2}\right) = 0$$

$$-\frac{1}{2}(-2,+2) \times \begin{pmatrix} 1 & 0 \\ 0 & 0,5626 \end{pmatrix} \times \begin{pmatrix} -2 \\ 2 \end{pmatrix} + (x_1, x_2) \times \begin{pmatrix} 1 & 0 \\ 0 & 0,5626 \end{pmatrix} \times \begin{pmatrix} -2 \\ 2 \end{pmatrix} = 5.56 - 2x_1 + 3.56 x_2$$

Exemple 2

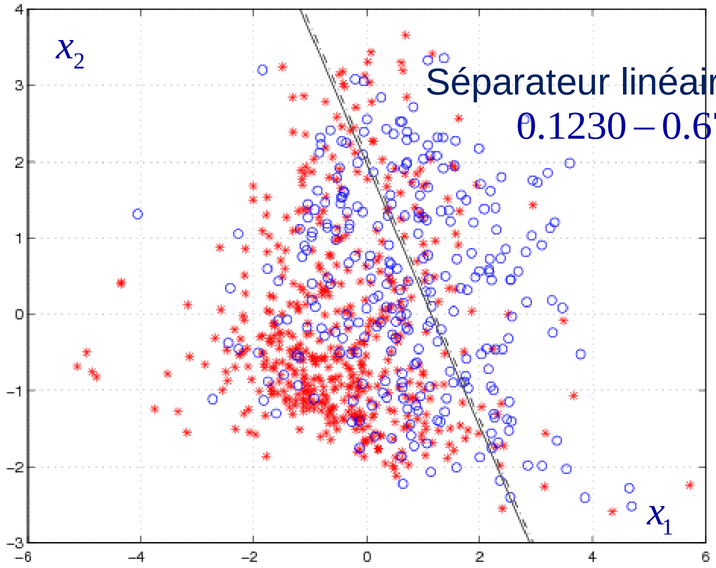
$$\pi_1 = 0.651$$

$$\pi_2 = 0.349$$

$$\mu_1 = (-0.4035, -0.1935)^t$$

$$\mu_2 = (0.7528, 0.3611)^t$$

$$\Sigma = \begin{pmatrix} 1.7925 & -0.1461 \\ -0.1461 & 1,6634 \end{pmatrix}$$



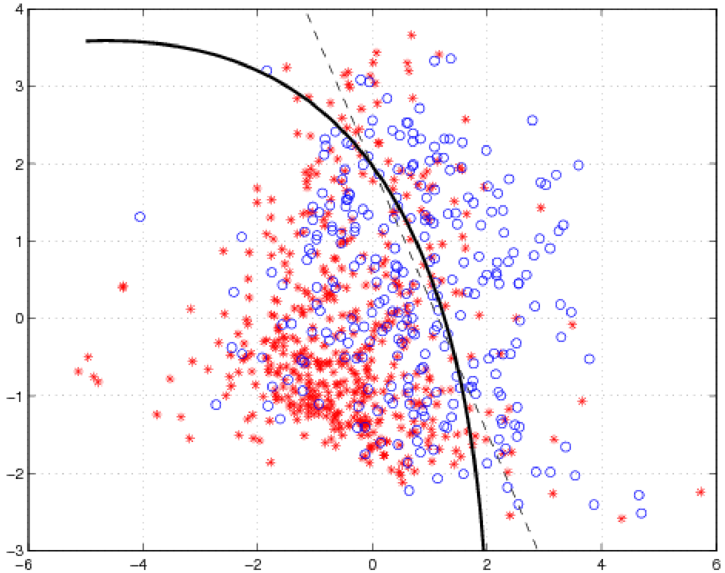
Séparateur linéaire

$$0.1230 - 0.6770 x_1 - 0.3928 x_2 = 0$$

Exemple 3: Séparateur quadratique

Le calcul de la fonction de séparation (cf.https://stats.stackexchange.com/questions/252800/calculate-the-decision-boundary-for-quadratic-discriminant-analysis-qda

- Prior probabilities: $\hat{\pi}_1 = 0.651$, $\hat{\pi}_2 = 0.349$.
- $\hat{\mu}_1 = (-0.4035, -0.1935)^T$, $\hat{\mu}_2 = (0.7528, 0.3611)^T$.
- $\hat{\Sigma}_1 = \begin{pmatrix} 1.6769 & -0.0461 \\ -0.0461 & 1.5964 \end{pmatrix}$
- $\hat{\Sigma}_2 = \begin{pmatrix} 2.0087 & -0.3330 \\ -0.3330 & 1.7887 \end{pmatrix}$



- Le jeu de données est constitué de variables catégorielles (mais pas que..)

Attribut (a_1) Attribut (a_p) Classification (**classe c**) → Variable catégorielle

Instance →

Salaire	impôts	étudiant	contrôle
< 30	< 20 %	oui	non
30-50	< 20 %	non	oui
30-50	< 20 %	oui	oui
30-50	> 20 %	non	non
> 50	>20 %	non	oui

Ensemble d'apprentissage

35	6%	oui	?

L'apprentissage Bayésien permet de faire des prédictions en se basant sur des probabilités

- Pour une instance donnée (un individu – ligne du tableau), quelle sera la classification la plus probable (contrôle oui / non) en fonction de l'ensemble d'apprentissage (données d'entrainement)

Le classifier Bayésien naif

*On suppose que les descripteurs sont **deux à deux indépendants** conditionnellement aux valeurs de la variable à prédire*

- $P(c_k)$ = proportion d'instances de la classe contrôle (k = oui / k = non)

  P(contrôle = oui) = 3/5

  P(contrôle = non) = 2/5

contrôle
non
oui
oui
non
oui

- $P(a_1,\ldots,a_n)$ = proportion d'instances pour les attributs a1… ap

Salaire	impôts	étudiant	contrôle
< 30	< 20 %	oui	non
30-50	< 20 %	non	oui
30-50	< 20 %	oui	oui
30-50	> 20 %	non	non
> 50	>20 %	non	oui

$$P(a_1 \cap a_2 \cap \ldots \cap a_p = P(a_1) \cap P(a_2) \cap \ldots \cap P(a_p) = \prod_{i=1}^{p} P(a_i)$$

si indépendance: Hypothèse forte

Ex: P(Salaire <30 et Impôt < 20 et Etudiant = Oui)

P(Salaire <30 ∩ Impôt < 20 ∩ Etudiant = Oui)

P(Salaire <30) ∩ P(Impôt < 20) ∩ P(Etudiant = Oui)

P(Salaire <30) × P(Impôt < 20) × P(Etudiant = Oui)

1/5 × 3/5 × 2/5

Salaire	impôts	étudiant	contrôle
< 30	< 20 %	oui	non
30-50	< 20 %	non	oui
30-50	< 20 %	oui	oui
30-50	> 20 %	non	non
> 50	>20 %	non	oui

- Probabilité d'observer une instance **connaissant a priori** la classe d'appartenance

Il s'agit d'une probabilité conditionnelle telle que:

$P(a_1,\ldots,a_n / c_k)$ = nombre de fois où l'on rencontre $(a_1,\ldots,a_n)$ dans les instances de la classe c_k

Cette probabilité est appelée la **vraisemblance** (a priori)

Si tous les attributs sont indépendants alors, on peut décomposer cette probabilités comme suit:

$$P(a_1 \cap a_2 \cap \ldots \cap a_p / c_k) = P(a_1 / c_k) \cap P(a_2 / c_k) \cap \ldots \cap P(a_p / c_k) = \prod_{i=1}^{p} P(a_i / c_k) \longrightarrow \prod_{i=1}^{p} \frac{P(a_i \cap c_k)}{P(c_k)}$$

Ex: P(Salaire <30 et Impôt < 20 et Etudiant = Oui / contrôle = Oui)

P(Salaire = 30-50 / contrôle = Oui) ∩ P(Impôt < 20 / contrôle = Oui) ∩ P(Etudiant = Oui/ contrôle = Oui)

P(Salaire = 30-50 / contrôle = Oui) x P(Impôt < 20 / contrôle = Oui) x P(Etudiant = Oui/ contrôle = Oui)

$$\frac{P(\text{Salaire} = 30\text{-}50 \cap \text{contrôle} = \text{Oui})}{P(\text{contrôle} = \text{Oui})} = \frac{2/5}{3/5} \qquad = \frac{2/5}{3/5} \qquad = \frac{1/5}{3/5}$$

Salaire	impôts	étudiant	contrôle
< 30	< 20 %	oui	non
30-50	< 20 %	non	oui
30-50	< 20 %	oui	oui
30-50	> 20 %	non	non
> 50	>20 %	non	oui

		Etudiant		
		Oui	Non	tot
Contrôle	Oui	1	2	3
	Non	1	1	2
	Tot	2	3	5

Etudiant = Oui et Contrôle = Oui sont dépendants

- On sait donc calculer les probabilités de chaque instance connaissant a priori la classe d'appartenance $\prod_{i=1}^{p} \frac{P(a_i \cap c_k)}{P(c_k)}$

On cherche à calculer la classe d'appartenance en fonction des instances

35	6%	oui	?

probabilité d'appartenance à une classe

connaissant les valeurs des instances

On cherche à calculer la probabilité d'appartenance à une classe connaissant les valeurs des instances et donc

$$P(c_k / a_1, a_2, \ldots, a_p) = P(c_k / a_1 \cap a_2 \cap \ldots \cap a_p)$$

Pour y parvenir, on utilise le théorème de Bayes

$$\left.\begin{array}{l} P(E_1 / E_2) = \dfrac{P(E_1 \cap E_2)}{P(E_2)} \\ P(E_2 / E_1) = \dfrac{P(E_2 \cap E_1)}{P(E_1)} \end{array}\right\} P(E_1 / E_2) = \frac{P(E_2 / E_1) \times P(E_1)}{P(E_2)}$$

Vraisemblance

$$P(c_k / a_1, a_2, \ldots, a_p) = \frac{P(a_1, a_2, \ldots, a_p / c_k) \times P(c_k)}{P(a_1, a_2, \ldots, a_p)}$$

Probabilité à *priori*

Probabilité à *postériori*

Pour chaque instance, on calcule les probabilités a posteriori pour le contrôle = Oui et le contrôle = Non

P(contrôle = Oui / salaire = 30-50 et impôt <20% et étudiants = Oui)

P(contrôle = Non / salaire = 30-50 et impôt <20% et étudiants = Oui)

$$P(c_k / a_1, a_2, \ldots, a_p) = \frac{P(a_1, a_2, \ldots, a_p / c_k) \times P(c_k)}{P(a_1, a_2, \ldots, a_p)}$$

$$\text{P(contrôle = Oui / salaire = 30-50 et impôt <20\% et étudiants = Oui)} = \frac{\text{P(salaire = 30-50 et impôt <20\% et étudiants = Oui /contrôle = Oui /)} \times \text{P(contrôle = Oui)}}{\text{P(salaire = 30-50 et impôt <20\% et étudiants = Oui)}}$$

$$\text{P(contrôle = Non / salaire = 30-50 et impôt <20\% et étudiants = Oui)} = \frac{\text{P(salaire = 30-50 et impôt <20\% et étudiants = Oui / contrôle = Non)} \times \text{P(contrôle = Non)}}{\text{P(salaire = 30-50 et impôt <20\% et étudiants = Oui)}}$$

- On choisit la classe la plus probable (Probabilité maximum a postériori)

$$h_{MAP} = \arg\max_{c_k} \left(\frac{P(a_1, a_2, ..., a_n / c_k) * P(c_k)}{P(a_1, a_2, ..., a_n)} \right)$$

$P(a_1, a_2, ..., a_p) = Cte$ (Indépendante de la variable à prédire)

On sélection la probabilité telle que $h_{MAP} = \arg\max_{c_k} \left(P(a_1, a_2, ..., a_n / c_k) * P(c_k) \right)$

- Retour à l'exemple

P(salaire = 30-50 et impôt <20% et étudiants = Oui /) x P(contrôle = Oui)

1 P(salaire = 30-50 et impôt <20% et étudiants = Oui /contrôle = Oui /)

2 P(contrôle = Oui) = 3/5

1 $\rightarrow$ P(Salaire = 30-50 / contrôle = Oui) x P(Impôt < 20 / contrôle = Oui) x P(Etudiant = Oui/ contrôle = Oui)

		Salaire			
		< 30	30-50	> 50	contrôle
contôle	Oui	0	2	1	3
	Non	1	1	0	2
	Salaire	1	3	1	5

2/3

		Impot		
		<20	>20	contrôle
contôle	Oui	2	1	3
	Non	1	1	2
	impot	3	2	5

2/3

		Etudiant		
		Oui	Non	contrôle
contôle	Oui	1	2	3
	Non	1	1	2
	impot	2	3	5

1/3

P(Salaire = 30-50 / contrôle = Oui) x P(Impôt < 20 / contrôle = Oui) x P(Etudiant = Oui/ contrôle = Oui)

=

2/3 x 2/3 x 1/3 = 0.148

P(contrôle = Oui / salaire = 30-50 et impôt <20% et étudiants = Oui) x P(contrôle = Oui) = 0.148 * 0.6 = 0.088

P(contrôle = Non / salaire = 30-50 et impôt <20% et étudiants = Oui) × P(contrôle = Non)

1

2 P(contrôle = Non) = 2/5

P(Salaire = 30-50 / contrôle = Non) x P(Impôt < 20 / contrôle = Non) x P(Etudiant = Oui/ contrôle = Non)

		Salaire			
		< 30	30-50	> 50	contrôle
contôle	Oui	0	2	1	3
	Non	1	1	0	2
	Salaire	1	3	1	5

1/2

		Impot		
		<20	>20	contrôle
contôle	Oui	2	1	3
	Non	1	1	2
	impot	3	2	5

1/2

		Etudiant		
		Oui	Non	contrôle
contôle	Oui	1	2	3
	Non	1	1	2
	impot	2	3	5

1/2

P(Salaire = 30-50 / contrôle = Non) x P(Impôt < 20 / contrôle = Non) x P(Etudiant = Oui/ contrôle = Non)

=

1/2 x 1/2 x 1/ 2 = 0.125

P(contrôle = Non / salaire = 30-50 et impôt <20% et étudiants = Oui)× P(contrôle = Non) = 0.125 x 0.4 = 0.05

$$P(a_1, a_2, \ldots, a_p \,/\, c = Oui) > P(a_1, a_2, \ldots, a_n \,/\, c = Non)$$

On effectue un contrôle fiscal

Correctif sur le calcul de la vraisemblance

En pratique, on utilise un facteur correctif (qui permet de « lisser » les estimations) ce qui évite d'obtenir des probabilités conditionnelles nulles et par conséquent des vraisemblances nulles

		Salaire			
		< 30	30-50	> 50	contrôle
contôle	Oui	0	2	1	3
	Non	1	1	0	2
	Salaire	1	3	1	5

$$P(Salaire < 30 \,/\, Controle = Oui) = \frac{P(Salaire < 30 \cap Controle = Oui)}{P(Controle = Oui)}$$

$$P(Salaire < 30 \cap Controle = Oui) = 0!$$

- Pour palier au problème de probabilité conditionnelle nulle, on utilise l'estimateur Laplacien des probabilités

n_{kl}

		Salaire			
		< 30	30-50	> 50	contrôle
contôle	Oui	0	2	1	3
	Non	1	1	0	2
	Salaire	1	3	1	5

$\longrightarrow n_k$

$$P(a_i / c_k) = \frac{n_{kl}}{n_k} \longrightarrow P(a_i / c_k) = \frac{n_{kl} + m}{n_k + mk}$$

m = 1
k = nombre de niveaux de la variable à prédire

- Nouvelle estimation $P(Salaire < 30 / Controle = Oui) = \frac{P(Salaire < 30 \cap Controle = Oui)}{P(Controle = Oui)} = \frac{1}{(3+2)}$
- On applique cette modification pour le calcul de toutes les probabilités conditionnelles
- Il n'y a pas de modification du critère de décision (h_{MAP})

Données manquantes

Les données manquantes sont traitées de façon satisfaisante par le classifieur. Pour l'estimation des probabilités, ces valeurs sont tout simplement ingnorées

Salaire	impôts	étudiant	contrôle
< 30	< 20 %	oui	non
30-50	< 20 %	non	oui
?	< 20 %	oui	oui
30-50	> 20 %	non	non
> 50	>20 %	non	oui

P(Salaire = 30-50 / contrôle = Oui) $\cap$ P(Impôt < 20 / contrôle = Oui) $\cap$ P(Etudiant = Oui/ contrôle = Oui)

1 x P(Impôt < 20 / contrôle = Oui) x P(Etudiant = Oui/ contrôle = Oui)

Variables prédictives numériques

Salaire	impôts	étudiant	contrôle
18	< 20 %	oui	non
22	< 20 %	oui	non
35	< 20 %	non	oui
40	< 20 %	oui	oui
44	> 20 %	non	non
56	>20	non	oui
58	>20 %	non	oui

→ Calcul des probabilités conditionnelles pour la variable salaire

$$P(Salaire / Contrôle = Oui)$$

$$P(Salaire / Contrôle = Non)$$

- Pour calculer les probabilités conditionnelles d'une variable numérique *a*, on suppose que la distribution de cette variable suit **une distribution normale** pour chaque niveau de facteur de la variable c_k

$$a_{c_k} \rightarrow \mathbb{N}(\mu_{a,c_k}, \sigma^2_{a,c_k})$$

Densité de probabilité

$$F_{p,c_k}(x) = \frac{1}{\sigma_{p,c_k}\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{x-\mu_{p,c_k}}{\sigma_{p,c_k}}\right)^2 dx$$

- Concernant notre exemple : pour la variable salaire

$$Salaire_{contrôle=Oui} \rightarrow \mathbb{N}(\mu_{controle=Oui}, \sigma^2_{controle=Oui})$$

$$Salaire_{contrôle=Non} \rightarrow \mathbb{N}(\mu_{controle=Non}, \sigma^2_{controle=Non})$$

$$f_{contrôle=Oui}(x) = \frac{1}{\sigma_{contôle=Non}\sqrt{2\pi}} \exp\left(\frac{x-\mu_{contôle=Oui}}{\sigma_{contôle=Oui}}\right)^2$$

$$f_{contrôle=Non}(x) = \frac{1}{\sigma_{contôle=Non}\sqrt{2\pi}} \exp\left(\frac{x-\mu_{contôle=Non}}{\sigma_{contôle=Non}}\right)^2$$

Probabilité conditionnelle pour x = 38 ?

	Contrôle	
	Oui	Non
	35	18
	40	22
	56	44
	58	
Moyenne	47,25	28,00
ecart type	11,47	14,00

$$f_{contrôle=Non} = P(Salaire = 38 / contrôle = Non) = 0.0220$$

$$f_{contrôle=Oui} = P(Salaire = 38 / contrôle = Oui) = 0.0251$$

Le fonction *hmap* est passée en logarithme. On cherche donc à minimiser le log de la vraisemblance à posteriori

$$h_{MAP} = \arg\max_{c_k} \left(\frac{P(a_1, a_2, ..., a_n / c_k) * P(c_k)}{P(a_1, a_2, ..., a_n)} \right)$$

$$\log(h_{map}) = \arg\max \Big(\log\big(P\big(a_1, a_2, ..., a_n / c_k\big)\big) + \log\big(P(c_k) - \log\big(P(a_1, a_2, ..., a_n)\big)\Big)$$

Sachant que $\log\big(P(a_1, a_2, ..., a_n)\big) = cte$

$$\log(h_{map}) = \arg\max \Big(\log\big(P\big(a_1, a_2, ..., a_n / c_k\big)\big) + \log\big(P(c_k)\big)\Big)$$

$$\boxed{\log(h_{map}) = \arg\max \left(\log\left(\sum_{i=1}^{n} P(a_i / c_k) \right) + \log\big(P(c_k)\big) \right)}$$

D’un point de vue calculatoire, le produit de nombreuses probabilités (toutes inférieures à 1) provoque rapidement des débordements de capacité de la mémoire même en utilisant des flottants à double précision. Le passage en log permet d’éviter ce problème.

- Présente de bonnes performances (comparables à d’autre techniques plus « sophistiquées »)
- Facile à programmer, rapidité de calcul et capacité à traiter de très grandes bases
- Très robuste
- Règles d’affectations non interprétables et non explicites(comme d’autres algos…) car on ne peut pas estimer comme chaque variable « pèse » sur la décision
- Les données numériques ne sont pas toujours distribuées selon un loi normale. Dans ce cas, on peut toujours utiliser des transformations (Cox-Box)