Apprentissage Automatique Numérique

—

Naive Bayes

Loïc BARRAULT

loic.barrault@univ-lemans.fr
Laboratoire d'Informatique de l'Université du Maine

14 septembre 2021

Classification Automatique

Autre terme : reconnaissance de formes

Problème classique

- Tâche : distinguer plusieurs objets
- Association d'une catégorie (classe) à un objet inconnu
- Généralement les objets à classer sont représentés par des données numériques
- On distingue deux types d'approches :
  - Classification supervisée
  - Classification non-supervisée
- Dans ce cours : principalement la classification supervisée

Classification supervisée

Caractéristiques :

- Le nombre et le type des classes sont fixes et connus d'avance (*rejet* possible si aucune classe ne convient)
- On dispose d'exemples typiques, chacun associé à la classe souhaitée

Exemples d'application :

- Reconnaissance Optique de Caractères (OCR)
  - écriture manuscrite : chèques, adresses postales, ...
  - écriture tapuscrite : livres, revues, magazines
- Reconnaissance Automatique de la Parole (ASR)
- Photo : détection de sourire, de visage
- Météo : classification d'images satellites
- ...

Classification non-supervisée

Caractéristiques :

- Aucune information sur le **nombre** et le **type** des classes
- On dispose juste d'un jeu de données
- 2 étapes classiques :
  - **regrouper** les données (**caractéristiques communes**)
  - **identifier** ces regroupements

Quelques questions se posent et s'imposent :

→ Y a-t-il des ressemblances entre les individus ?

→ Quels critères pour définir cette ressemblance ?

→ Est-il pertinent de faire plusieurs groupes ?

→ Comment déterminer le nombre de groupes ?

→ Comment estimer la qualité de la classification ?

→ Généralement, on sait répondre après coup ... !

Combinatoire :

- Objectif : partitionner les exemples de manière optimale en fonction de certains critères
- Question : peut-on explorer toutes les solutions possibles et choisir la meilleure ?
  - Pour séparer un ensemble E composé de n exemples en K classes :
  - Nombre de partitions possibles de E en K classes (nombre de Stirling de première espèce) : $s(n, K) \sim K^n/K!$
  - Nombre total de partitions (nombre de Bell) :

$$B_n = \sum_{k=1}^{n} s(n, k) = \frac{1}{e} \sum_{k \geq 1} \frac{k^n}{k!}$$

- Stratégies itératives : exploration d'un sous-ensemble des solutions

Exemples d'application :

- Analyse de données en général
- Classifier les clients dans un supermarché en fonction de leurs achats
- Identifier des groupes à risque pour une assurance
- Prise de décision : faut-il vendre ou acheter des actions ?
- ...

Représentation des données

- Tout individu est représenté par des valeurs numériques
  - obtenues automatiquement
  - permettant de le caractériser

- Tri automatique de poissons :
  - Longueur, diamètre, poids, ...
- Reconnaissance d'écriture :
  - image de taille 16×16 avec des valeurs de gris
  - nombre de traits dans l'image, contours, ...?
- Classification d'image satellite :
  - nombre, taille et couleur des zones, taille des zones homogènes contiguës, ...
- Détection de sourire :
  - traitement d'image $\rightarrow$ caractéristiques pertinentes

Représentation des données

- Quelles données ?
  - propriétés caractéristiques de l'individu

  $\rightarrow$ binaires, discrètes, continues

  $\rightarrow$ quantitatives : associées à une valeur numérique

  $\rightarrow$ qualitatives : il faut les représenter numériquement !
- Calcul de la représentation numérique peut être un problème compliqué
  - ex : traitement d'image complexe et lent
- Dualité :
  - Codage sophistiqué $\rightarrow$ la classification est simplifiée
  - Codage simple $\rightarrow$ la classification peut être plus complexe

$\Rightarrow$ Le bon choix du codage est très important

Le classifieur bayésien

Problème (selon Duda & Hardt)

- Cas pratique : sur un bateau de pêche, un tapis roulant fait défiler des poissons

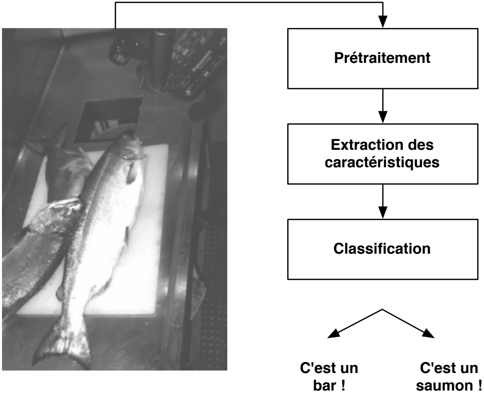

→ Comment séparer automatiquement bars et saumons ?

- Il faut un ou plusieurs critères de distinction

→ Consulter un **expert** (pisciculteur) :
  - largeur, longueur, couleur, nombre de nageoires, poids, ...

- Prise d'une photo du poisson
- Codage : calcul de ces caractéristiques **automatiquement**
  - traitement d'image ⇒ vecteur $\mathbf{x} \in \mathbb{R}^n$

Le classifieur bayésien



Rappel apprentissage supervisé

- À notre disposition : des images de bars et de saumons

→ corpus d'entraînement

Le classifieur bayésien

- Comment peut-on évaluer la décision ?

$\rightarrow$ nombre de mauvaises classifications

- Pour les poissons : chaque erreur a un coût identique
- Mais pour détecteur de faux billets : rejeter un vrai billet est moins grave que d'accepter un faux billet
- Généralisation : associer un coût à chaque décision

$\Rightarrow$ Trouver la règle de décision qui minimise le coût total

Le classifieur bayésien

Une première approche

- On sait qu'il y a beaucoup plus de saumons que de bars sur le tapis
- En absence d'autres informations, il est raisonnable de toujours décider pour la classe la plus probable
- On peut obtenir ces probabilités *a priori* en comptant le nombre de bars et de saumons dans une période de temps

$$P(\omega_1) = \frac{n_{bars}}{n_{bars} + n_{saum}} \qquad P(\omega_2) = \frac{n_{saum}}{n_{bars} + n_{saum}}$$

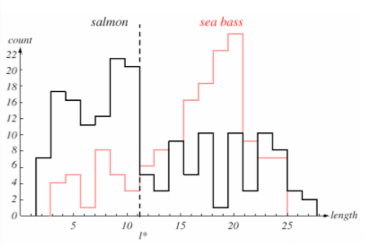
$$P(\omega_1) + P(\omega_2) = 1$$

- Il y a des tâches pour lesquelles le déséquilibre est plus prononcé (détecteur de faux billets)

Le classifieur bayésien

Une meilleure approche

- Comment utiliser les informations sur chaque poisson ?
- Les bars sont **généralement** plus longs que les saumons



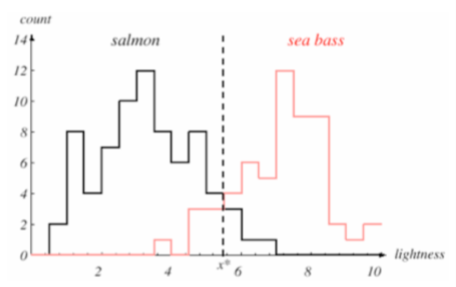
- Quel seuil appliquer pour faire la séparation ?
- Statistiques : $P(l|\omega_1)$ *et* $P(l|\omega_2)$

→ Chevauchement important : la taille seule n'est pas assez discriminante

Le classifieur bayésien

Autre caractéristique

- Les bars sont **généralement** plus lumineux que les saumons



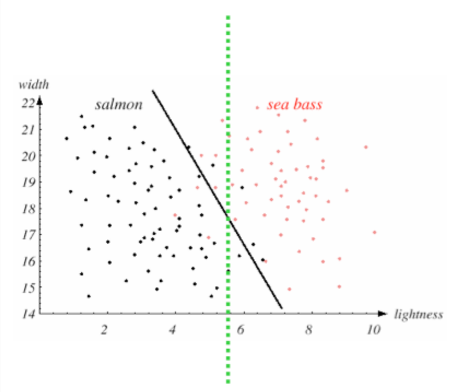
- Chevauchement moins important

$\rightarrow$ critère plus discriminant que la taille

- Statistiques : $P(x|\omega_1)$ *et* $P(x|\omega_2)$

Le classifieur bayésien

Combinaison des caractéristiques



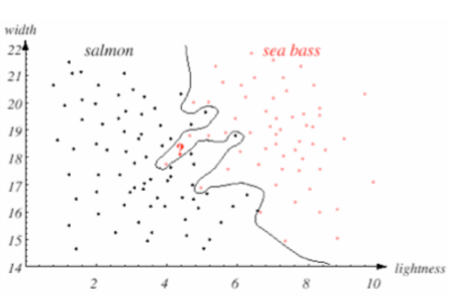
- Le seuil devient une courbe

$\rightarrow$ droite qui minimise le nombre d'erreur

Le classifieur bayésien

Division de l'espace

- Faut-il chercher le modèle qui explique le mieux les données d'apprentissage ?

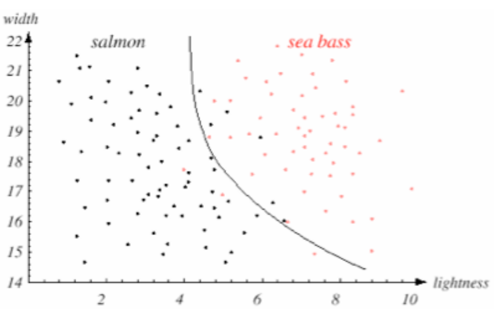


- Erreur = 0 sur le corpus d'entrainement
- Qu'en sera-t-il pour les nouveaux test ?

→ Il faut penser à la **généralisation**

Le classifieur bayésien

Un meilleur compromis ?



- Erreur plus grande sur le corpus d'entrainement
- Mais pouvoir de généralisation semble plus grand

→ manière d''éliminer les exemples confus et éviter le **sur-apprentissage**

Le classifieur bayésien

Dimension pour l'espace de représentation

- Faut-il ajouter toutes les caractéristiques imaginables ?
- Certaines peuvent ajouter plus de bruit que d'information
- Attention : redondance et/ou corrélation entre les caractéristiques
- Compromis entre le nombre de paramètres et le nombre d'exemples disponibles pour estimer ces paramètres

$\rightarrow$ **Fléau de la dimension** / Curse of dimensionality

Le classifieur bayésien

Partition des données

corpus d'**apprentissage** ou d'**entraînement** permet d'estimer les paramètres des modèles (ex. calcul d'une moyenne)

corpus de **développement** sert à prendre des décisions conceptuelles : quel est le meilleur modèle ? quels sont les meilleurs paramètres ?

corpus de **test** évaluation finale des performances du système

La règle de décision bayésienne

On a vu que ...

- À défaut d'autre information : $p(\omega_i) \rightarrow$ probabilité ***a priori***
- Avec 1 ou plusieurs critères : $p(x|\omega_i) \rightarrow$ ***vraisemblance***

$\rightarrow$ que vaut la vraisemblance quand l'***a priori*** est faible ?

- Ex : seul 1 poisson sur 100 est un saumon

La règle de décision bayésienne

On a vu que ...

- À défaut d'autre information : $p(\omega_i) \rightarrow$ probabilité ***a priori***
- Avec 1 ou plusieurs critères : $p(x|\omega_i) \rightarrow$ ***vraisemblance***

$\rightarrow$ que vaut la vraisemblance quand l'***a priori*** est faible ?

- Ex : seul 1 poisson sur 100 est un saumon

$\rightarrow$ Le classifieur Bayésien tient compte de ces 2 facteurs

La règle de décision bayésienne

- On choisit la classe dont la ***probabilité a posteriori*** est supérieure à celles des autres classes :

→ Choisir la classe la plus probable :

$$\omega^* = \underset{\omega_i}{\text{argmax}}\, P(\omega_i|x)$$

→ Mais on ne sait pas calculer directement les $P(\omega_i|x)$

La règle de décision bayésienne

- Règle de Bayes : $P(x|\omega_i)P(\omega_i) = P(\omega_i|x)P(x)$

$$\omega^* = \underset{\omega_i}{\operatorname{argmax}}\, P(\omega_i|x)$$
$$\omega^* = \underset{\omega_i}{\operatorname{argmax}}\, \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$$

- $P(\omega_i)$ probabilité ***a priori***
- $P(x|\omega_i)$ **densité de probabilité** de x pour la classe ω_i
- $P(\omega_i|x)$ probabilité ***a posteriori***
- Remarque :

$$P(x) = \sum_i P(x|\omega_i)P(\omega_i)$$

La règle de décision bayésienne

- Règle de Bayes :

$$\omega^* = \underset{\omega_i}{\operatorname{argmax}} \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$$

- $P(x)$ est constante pour toutes les classes ω_i
- Simplification finale

$$\omega^* = \underset{\omega_i}{\operatorname{argmax}}\, P(x|\omega_i)P(\omega_i)$$

Notion de l'Erreur

Formalisation

- soit $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_j \ldots \alpha_C\}$ l'ensemble des actions possibles

$\rightarrow$ en général : attribuer l'étiquette ω_j

- Soit λ_{ij} le coût engendré par l'action α_i lorsque l'objet appartient effectivement à la classe ω_j
- Cas particulier :

$$\lambda_{ij} = \begin{cases} 1 & \text{si } i \neq j \\ 0 & \text{sinon} \end{cases}$$

Le Risque

- Le risque associé à chaque action est :

$$R(\alpha_i|x) = \sum_j \lambda_{ij} P(\omega_j|x)$$

- Minimiser le risque, revient à prendre, pour chaque observation x, la décision qui minimise le risque conditionnel :

$$R(\omega_{i^*}|x) < R(\omega_i|x) \qquad \forall \mathrm{i} \neq \mathrm{i}^*$$

La règle de décision bayésienne

- Théorème : la règle de décision bayésienne est la règle de risque minimal
- Preuve : soit f_B le classifieur de Bayes et f un classifieur quelconque. ω_b et ω les classes proposées par ces 2 classifieurs.

$$\begin{aligned}
P(\omega_B|x) \geq P(\omega|x) &\Rightarrow P(x, \omega_B) \geq P(x, \omega) \\
P(x, y \neq \omega_B) &= \left(\sum_x P(x, \omega_i)\right) - P(x, \omega_B) \\
&\leq \left(\sum_x P(x, \omega_i)\right) - P(x, \omega) \\
&\leq P(x, y \neq \omega) \\
\text{et donc } R(f_B) &\leq R(f)
\end{aligned}$$

Exemples concrets

- Détecteur de faux billets
- Deux classes :
  - ω_1 vrai billet, $P(\omega_1) = 0.999$
  - ω_2 faux billet, $P(\omega_2) = 0.001$
- Deux actions :
  - α_1 accepter le billet
  - α_2 refuser le billet

Exemples concrets

Matrice des coûts :

- $\lambda_{11} = \lambda(\alpha_1|\omega_1) = 1$€ accepter un vrai billet (test)
- $\lambda_{12} = \lambda(\alpha_1|\omega_2) = 101$€ accepter un faux billet (test + perte)
- $\lambda_{21} = \lambda(\alpha_2|\omega_1) = 11$€ refuser un vrai billet (test + préjudice commercial)
- $\lambda_{22} = \lambda(\alpha_2|\omega_2) = 1$€ refuser un faux billet (test)

$\Rightarrow$ Les coûts inégaux décalent la frontière de décision

Utilisation du classifier Bayésien

Principe

- Le problème d'apprentissage est résolu si on connaît les $P(\omega_i)$ et $P(\omega_i|x)$
- Ceci permettra de construire un classifieur dont la probabilité d'erreur est minimale

Estimation des probabilités

- Utiliser les données d'un ensemble d'apprentissage pour obtenir une estimation de ces probabilités

Estimation des Probabilités *a priori*

- Sans informations supplémentaires, on suppose que les classes sont équiprobables :

$$\hat{p}(\omega_i) = \frac{1}{C}$$

- On utilise un ensemble d'apprentissage représentatif pour estimer les probabilités *a priori* par fréquence relative :

$$\hat{p}(\omega_i) = \frac{n_i}{\sum_i n_i} = \frac{n_i}{n}$$

- Le corpus d'apprentissage doit avoir une ***taille suffisante***

Estimation des Probabilités $p(x|\omega_i)$

Méthodes paramétriques

- On suppose que les $p(x|\omega_i)$ ont une certaine forme analytique (p.ex. une distribution normale)
- On utilise le corpus d'apprentissage pour estimer les ***paramètres*** de cette forme

Méthodes non paramétriques

- On estime les $p(x|\omega_i)$ au point x en observant les données du corpus d'apprentissage dans le voisinage de x

$\rightarrow$ Ceci n'est pas traité dans ce cours

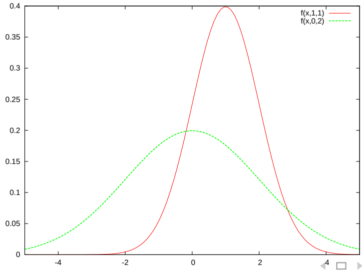
La Distribution Normale en 1D

- Aussi appelé Gaussienne
- Équation pour $d = 1$:

$$p(x) = \frac{1}{\sigma\sqrt{2\cdot\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- avec

$$\begin{aligned}\mu &= \text{moyenne}\\ \sigma &= \text{variance}\end{aligned}$$



La Distribution Normale en 2D

- Vecteur moyen :

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

- Matrice de covariance

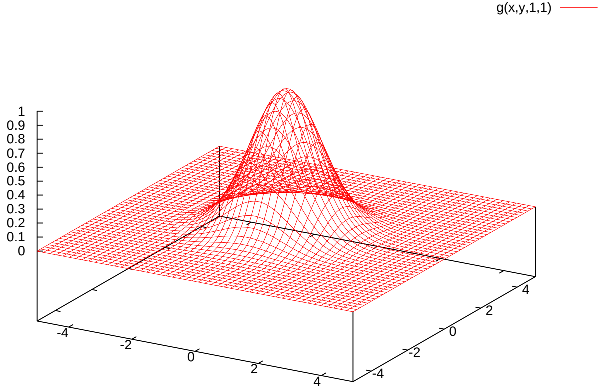
$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

- Équation pour $d = 2$:

$$p(\mathbf{x}) = \frac{1}{\sqrt{2 \cdot \pi} \|\Sigma\|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1} (\mathbf{x}-\mu)}$$

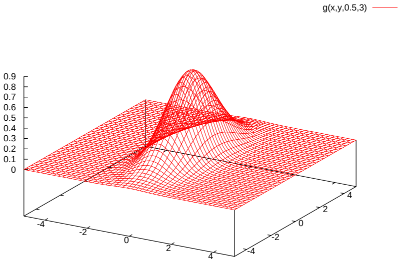
- Note : ceci est l'équation de la densité de probabilité, la valeur peut donc dépasser 1.

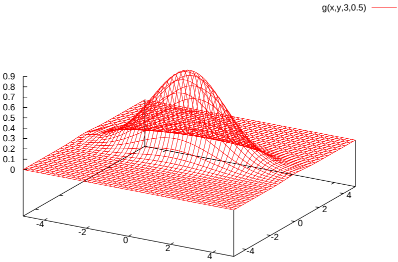
La Distribution Normale en 2D



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

La Distribution Normale en 2D

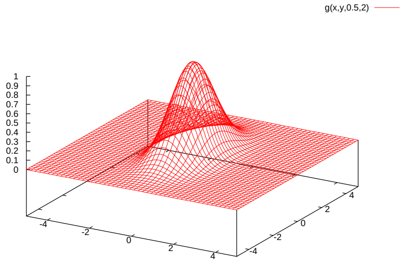


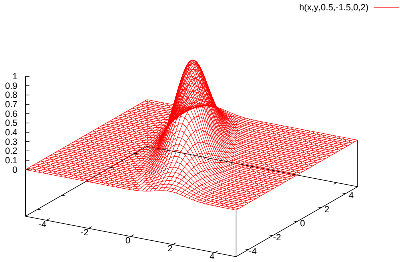


$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 3 \end{pmatrix}$$

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 0.5 \end{pmatrix}$$

La Distribution Normale en 2D





$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} 0.5 & -1.5 \\ 0 & 2 \end{pmatrix}$$

La Distribution Normale en $\mathbb{R}^d$

- Vecteur moyen :

$$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$$

- Matrice de covariance :

$$\Sigma = (\sigma_{ij})$$

- Équation pour $d = 2$:

$$p(x) = \frac{1}{\sqrt{2 \cdot \pi} \|\Sigma\|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^t \Sigma^{-1} (\mathbf{x}-\mu)}$$

  Note : ceci est l'équation de la densité de probabilité, la valeur peut donc dépasser 1.

- Facile à programmer/disponible en Matlab/Scilab, **python**

Estimation d'une Gaussienne

- Hypothèse : les $p(\mathbf{x}|\omega_i)$ suivent une loi Gaussienne
$$\mathcal{N}(\mu, \Sigma)$$
- Il faut estimer μ et Σ à partir des données d'apprentissage
- On peut monter que :

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}^{(k)} \\ \hat{\Sigma} &= \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}^{(k)} - \hat{\mu})^t (\mathbf{x}^{(k)} - \hat{\mu}) \end{aligned}$$

- Ces calculs sont fait séparément pour les exemples de chaque classe

Exemple applicatif (source : wikipedia)

- Soit le corpus d'entraînement suivant :

Sexe **S**	Taille **T** (cm)	Poids **P** (kg)	Pointure **Pt** (cm)
M	182	81.6	30
M	180	86.2	28
M	170	77.1	30
M	180	74.8	25
F	152	45.4	15
F	168	68.0	20
F	165	59.0	18
F	175	68.0	23

1. Calculer les probabilités ***a priori*** de chaque classe ω_i
2. les probabilités conditionnelles (**vraisemblance**) $p(x|\omega_i)$
3. les probabilités ***a posteriori*** $p(\omega_i|x)$

   (on omettra la **constante de normalisation** $p(x)$)

Exemple applicatif (source : wikipedia)

- On doit obtenir cela :

S	$\mu(T)$	$\sigma^2(T)$	$\mu(P)$	$\sigma^2(P)$	$\mu(Pt)$	$\sigma^2(Pt)$
M	178	2.93e+01	79.92	2.55e+01	28.25	5.58e+00
F	165	9.27e+01	60.1	1.14e+02	19.00	1.13e+01

- L'individu suivant est-il un homme ou une femme ?

Sexe	**Taille (cm)**	**Poids (kg)**	**Pointure (cm)**
inconnu	183	59	20

Exemple applicatif (source : wikipedia)

- Pour la classe "F" :
- ***a priori*** : P(F) $= \frac{4}{8} = 0.5$
- Vraisemblances : $\frac{1}{\sigma\sqrt{2\cdot\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  - $P(T = 183|F) = \frac{1}{9.63\sqrt{2\cdot\pi}} e^{-\frac{1}{2}\left(\frac{183-165}{9.63}\right)^2} = 7.21e-3$
  - $P(P = 59|F) = \frac{1}{10.7\sqrt{2\cdot\pi}} e^{-\frac{1}{2}\left(\frac{59-60.1}{10.7}\right)^2} = 3.72e-2$
  - $P(Pt = 20|F) = \frac{1}{3.36\sqrt{2\cdot\pi}} e^{-\frac{1}{2}\left(\frac{20-19}{3.36}\right)^2} = 1.13e-1$

  - ***a posteriori*** $P(F|x) =$
    $P(F) * P(T = 183|F) * P(P = 59|F) * P(Pt = 20|F)$
    $= 1.52e-5$

Exemple applicatif (source : wikipedia)

- La même chose pour la classe "M" :
- ***a priori*** : $\mathrm{P(M)} = \frac{4}{8} = 0.5$
- Vraisemblances : $\frac{1}{\sigma\sqrt{2 \cdot \pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
  - $P(T = 183|M) = \frac{1}{5.42\sqrt{2 \cdot \pi}} e^{-\frac{1}{2}\left(\frac{183-178}{5.42}\right)^2} = 4.81e-2$
  - $P(P = 59|M) = \frac{1}{5.05\sqrt{2 \cdot \pi}} e^{-\frac{1}{2}\left(\frac{59-79.92}{5.05}\right)^2} = 1.46e-5$
  - $P(Pt = 20|M) = \frac{1}{2.36\sqrt{2 \cdot \pi}} e^{-\frac{1}{2}\left(\frac{20-28.25}{2.36}\right)^2} = 3.81e-4$

  - ***a posteriori*** $P(M|x) =$
    $P(M) * P(T = 183|M) * P(P = 59|M) * P(Pt = 20|M)$
    $= 1.34e-10$