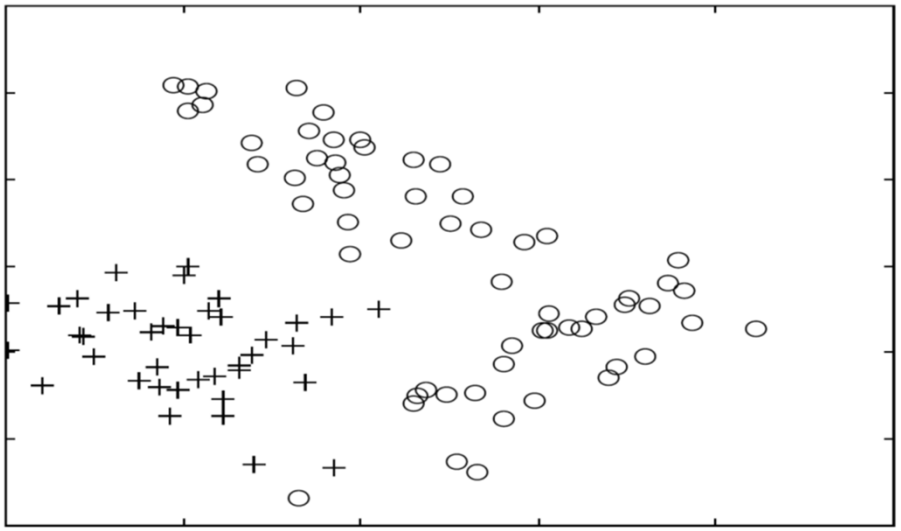
Classifieurs Paramétriques :

- **Classifieur Euclidien.**
- **Classifieur Quadratique.**
- **Classifieur Naïf Bayésien.**

Soit la représentation d'un objet quelconque au moyen d'un vecteur de caractéristiques $X = \left[x_1 \, x_2 \cdots x_d \right]^T$

Tous les vecteurs qui représentent l'ensemble des objets peuvent être positionnés dans l'espace Euclidien R^d, où ils correspondent chacun à un point.

Ces points peuvent alors être regroupés en amas, chacun de ces amas étant associé à une classe particulière.



Représentation d'objets appartenant à deux classes distinctes, dans un espace à deux dimensions.

Un classifieur doit être capable de modéliser au mieux les frontières qui séparent les classes les unes des autres. Cette modélisation fait appel à la notion de *fonction discriminante*, qui permet d'exprimer le critère de classification de la manière suivante:

" Assigner la classe ω_i à l'objet représenté par le vecteur X si, et seulement si, la valeur de la fonction discriminante de la classe ω_i est supérieure à celle de la fonction discriminante de n'importe quelle autre classe ω_j ".

Ou encore, sous forme mathématique:

$$X \in \omega_i \quad \Leftrightarrow \quad \Phi_i(X) \geq \Phi_j(X) \quad \forall j = 1,2,\ldots,C;\ j \neq i.$$

où $\Phi i(X)$ est appelé ***fonction discriminante*** de la classe ωi , et *C* est le nombre total de classes.

1- Classifieur Euclidien

Il s'agit probablement de l'un des plus simples classifieurs qui puissent être conçus.
La classe dont le vecteur de caractéristiques moyen est le plus proche, au sens de la distance Euclidienne, du vecteur de caractéristiques de l'objet à classifier est assignée à ce dernier. Les fonctions discriminantes utilisées sont donc de la forme suivante :

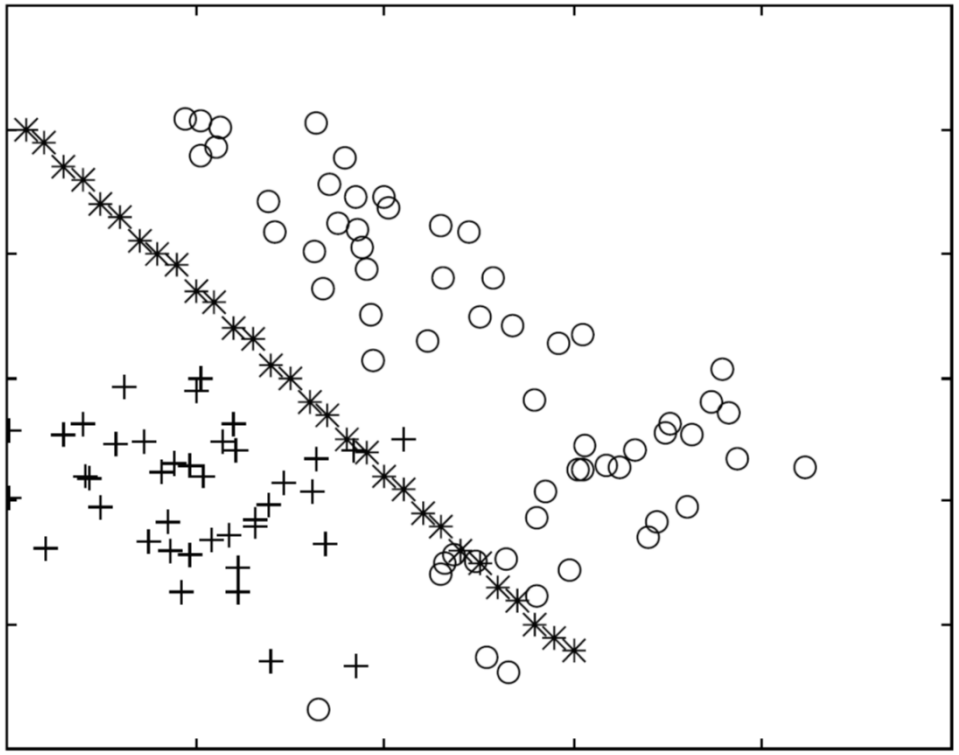
$$\Phi_i(X) = -\frac{1}{2}(X - \mathbf{M}_i)^T (X - \mathbf{M}_i)$$

où $\mathbf{M}_i = E\{X|\omega_i\}$ est le vecteur de caractéristiques moyen des éléments qui appartiennent à la classe ω_i , $E\{\cdot\}$ désignant l'opérateur d'espérance mathématique, et $(\cdot)^T$ celui de transposition.

Le terme quadratique $X^T X$ est indépendant de la classe de l'objet, et les fonctions discriminantes peuvent également s'écrire:

$$\Phi_i(X) = \mathbf{M}_i^T X - \frac{1}{2}\mathbf{M}_i^T \mathbf{M}_i$$

Les frontières qui séparent les classes dans l'espace R^d sont ici linéaires.



Frontière fournie par le classifieur Euclidien dans le cas d'un problème à deux classes.

En pratique, les vecteurs de caractéristiques moyens ne sont pas disponibles, et doivent être estimés à partir d'un ensemble fini de prototypes de chaque classe:

$$\hat{\mathbf{M}}_i = \frac{1}{N_i} \sum_{X_k \in \omega_i} X_k$$

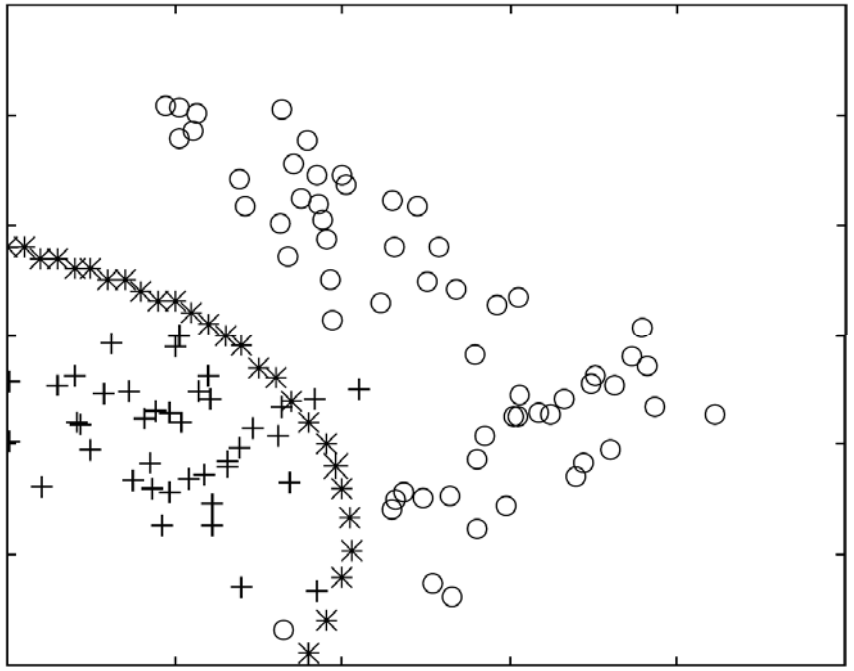
où N_i est le nombre total d'objets de classe ω_i qui sont à disposition, et X_k les vecteurs de caractéristiques qui représentent ces objets.

2- Classifieur Quadratique

Comme le nom l'indique, les frontières de décision fournies par ce modèle de classifieur sont quadratiques. Les fonctions discriminantes s'expriment :

$$\Phi_i(X) = -\frac{1}{2}(X - \mathbf{M}_i)^T \Sigma_i^{-1}(X - \mathbf{M}_i)$$

où $\Sigma_i = E\left\{(X - \mathbf{M}_i)(X - \mathbf{M}_i)^T | \omega_i\right\}$ est la matrice de covariance des vecteurs de caractéristiques de classe ω_i .



Frontière obtenue à l'aide du classifieur quadratique.

Tout comme les vecteurs de caractéristiques moyens de chaque classe, les matrices de covariance ne peuvent qu'être estimées à partir des objets disponibles.

$$\hat{\Sigma}_i = \frac{1}{N_i - 1} \sum_{X_k \in \omega_i} \left(X_k - \hat{\mathbf{M}}_i \right) \left(X_k - \hat{\mathbf{M}}_i \right)^T$$

Dans le cas particulier où les composantes des vecteurs de caractéristiques ne sont pas corrélées entre elles, les matrices de covariances expérimentales sont diagonales.
L'expression des fonctions discriminantes se réduit alors à:

$$\Phi_i(x) = -\frac{1}{2} \sum_{j=1}^{d} \frac{\left(x_j - \hat{\mu}_{ij} \right)^2}{\hat{\sigma}_{ij}^2}$$

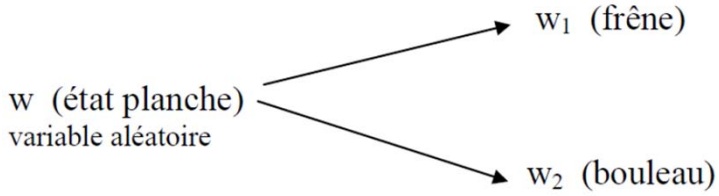
où $\hat{\mu}_{ij}$ et $\hat{\sigma}_{ij}^2$ représentent respectivement la moyenne et la variance expérimentales de la $j^{ème}$ composante du vecteur X, calculées sur les éléments de la classe ω_i.

3 - Classifieur Naif de Bayes

Introduction :

- On va considérer que chacune des classes est régie par un phénomène stochastique: chaque forme de la classe a une certaine probabilité de se produire et la loi de probabilité correspondante est gaussienne.

- On suppose que l'on s'intéresse à une entreprise de sciage de troncs d'arbres ayant la particularité de ne traiter que des frênes et des bouleaux. On va alors considérer l'état d'une planche (état = "frêne" ou "bouleau") comme une variable aléatoire w qui peut prendre deux valeurs w1 et w2.



Probabilité à priori d'une classe

Supposons que l'on connaît les quantités respectives de frêne et de bouleau qui rentrent dans la scierie, alors on connaît en fait les proportions de planches de frêne et de bouleau en sortie !

Supposons qu'on reçoit 2/3 de frêne et 1/3 de bouleau, alors on peut dire qu'en sortie deux planches sur trois en moyenne en sortie sont du frêne et une planche sur trois est du bouleau.

Si on doit maintenant décider la classe d'une planche quelconque qui sort et sur laquelle on n'a aucune information (on ne la voit pas , on ne la touche pas , on ne la sent pas ! ...) , on pourra valablement dire que c'est du frêne) et ainsi on minimise la probabilité de se tromper (puisqu'on a deux chances sur trois d'être dans le vrai).

Probabilité à priori d'une classe

En fait on peut dire qu'on dispose d'une information capitale qui est

- la probabilité à priori de w1: p(w1) = 2/3
- la probabilité à priori de w2: p(w2) = 1/3

Si on ne dispose même pas de la probabilité à priori de chacune des classes de formes recherchées alors en général on partagera la probabilité de façon égalitaire sur chacune des classes en concurrence.

Si on peut réaliser un apprentissage(échantillons représentatifs de chacune des classes fournis), dans ce cas la proportion de chacune des classes dans l'ensemble des échantillons permettra d'obtenir les probabilités à priori des classes.

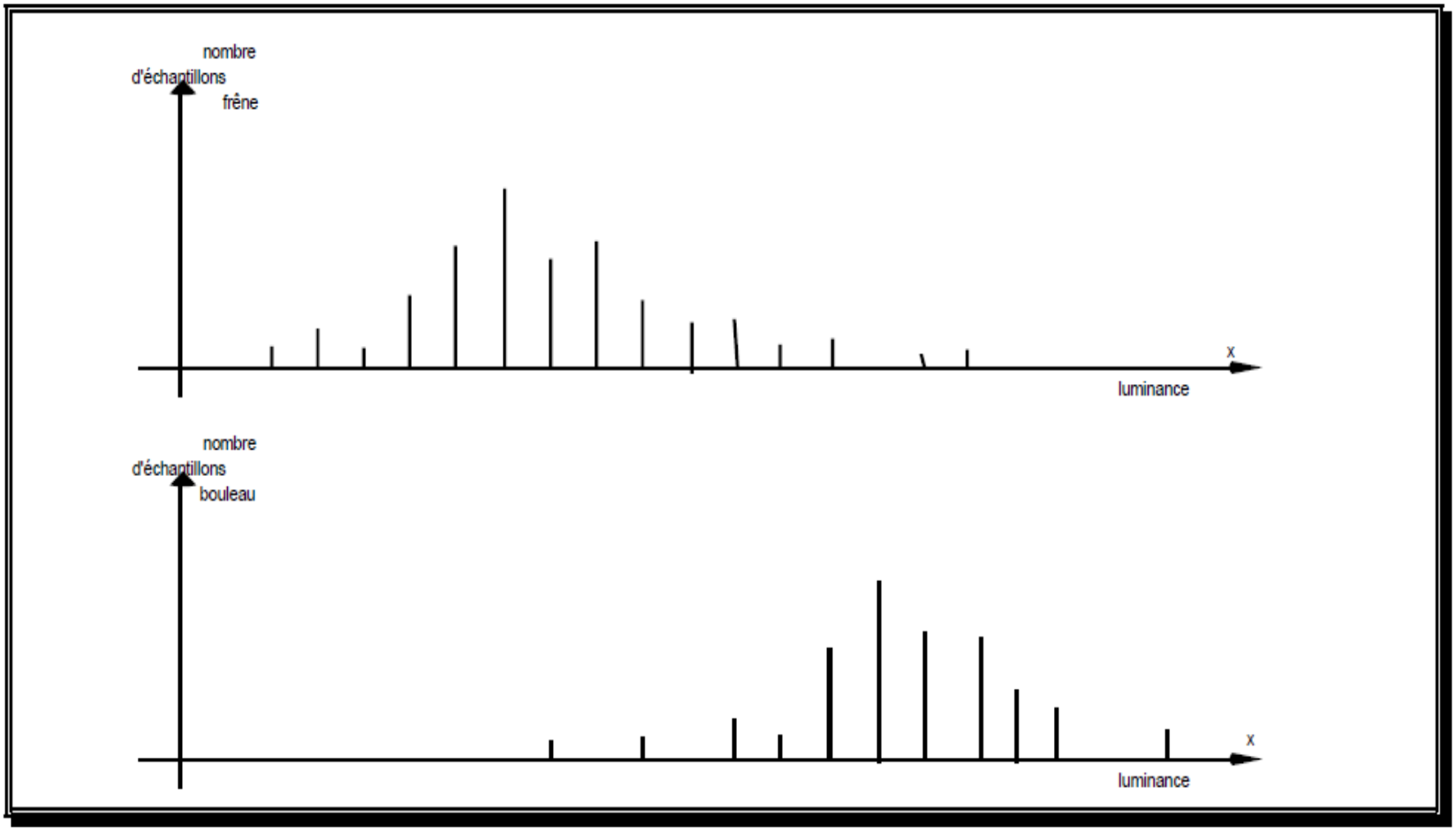
Loi de densité de probabilité d'une classe:

La scierie a installé une caméra N/B capable de saisir une image numérique de la surface de chaque planche qui sort et que cette image est traitée par un ordinateur qui peut fournir pour chaque image sa luminance moyenne. Ce paramètre étant considéré a priori comme relativement caractéristique de chaque essence de bois.

A partir d'échantillons (c'est à dire des planches qui sortent pour lesquelles on connaît la classe (w1 ou w2) et en même temps la luminance moyenne) on va chercher à estimer comment se distribue la luminance pour chacune des classes.

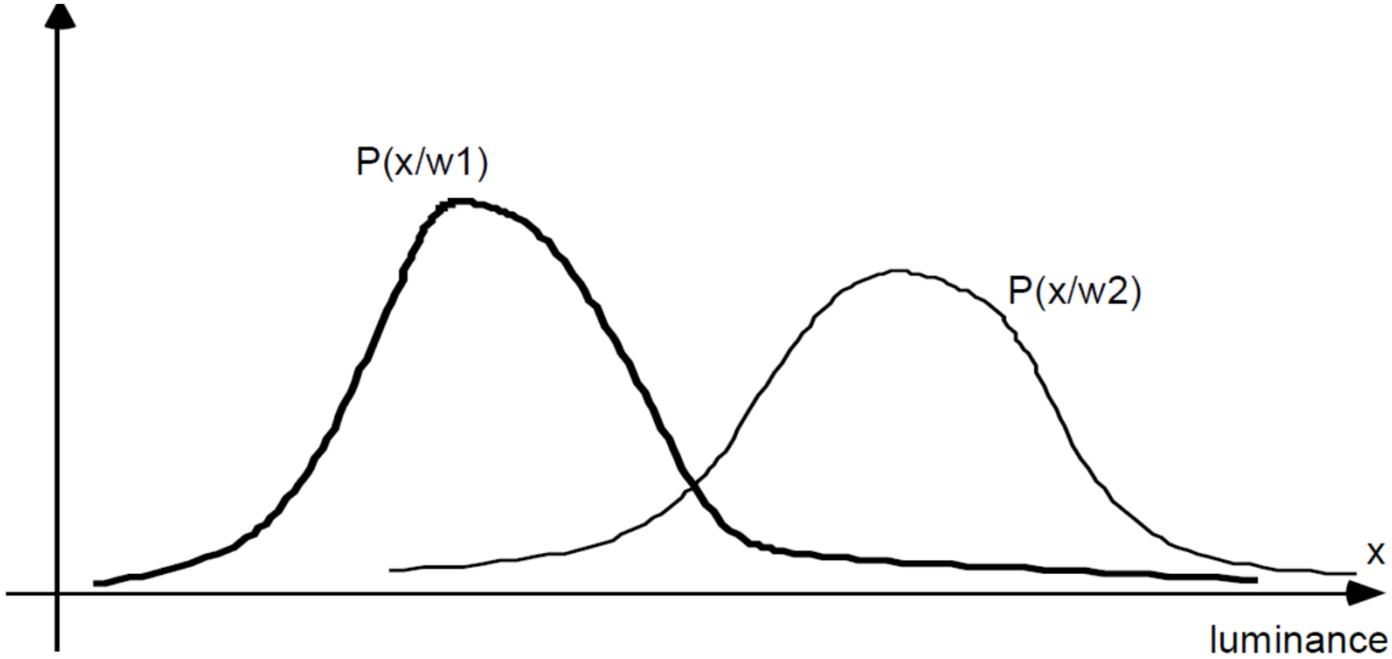
Loi de densité de probabilité d'une classe:

supposons avoir obtenu les histogrammes suivants:



Loi de densité de probabilité d'une classe:

Les fonctions de densité de probabilité de chacune des classes P(x/w1) et p(x/w2) représentées ci dessous:



Probabilité à posteriori d'une classe

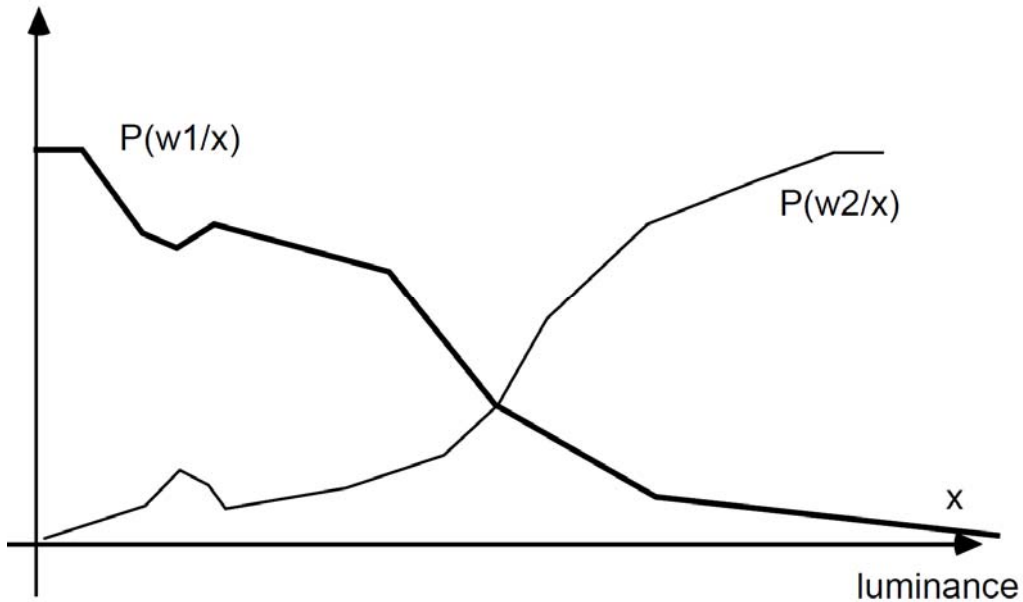
Règle de Bayes :

La règle de Bayes permet de calculer les probabilités à posteriori des classes , c'est à dire P(w1/x) et P(w2/x) à partir des probabilités à priori des classes P(w1) et P(w2) et des fonctions de densité des probabilités p(x/w1) et p(x/w2).

$$P(w_i/x) = \frac{P(x/w_i) \times P(w_i)}{P(x)} \quad \forall i = 1,2 \ \ avec \ \ P(x) = \sum_{j=1}^{2} P(x/w_j) \times P(w_j)$$

P(w1/x) : ayant obtenu une forme (décrite par une valeur x de luminance) P(w1/x) donne la probabilité d'avoir alors la classe w1.

Probabilité à posteriori d'une classe



La règle de décision Bayésienne est :

Ayant remarqué une forme inconnue x (luminance moyenne de la planche) on dira que la forme inconnue x est de la classe w1 **si P(w1/x) > P(w2/x)**, sinon elle est de la classe w2

Probabilité à posteriori d'une classe

En fonction de la décision prise pour classer une forme x on augmentera ou diminuera la probabilité d'erreur(probabilité de se tromper dans le choix de la classe d'affectation pour x)

Ainsi ayant une forme x on peut dire:

P(erreur/x) = P(w1/x) si on décide w2

P(w2/X) si on décide w1

On note que le bon choix est celui qui minimise la probabilité d'erreur.

L'erreur a en effet une probabilité minimale si on décide w1 et que P(w1/x) > P(w2/x) ou w2 et que P(w2/x) > P(w1/x).

Pertes et risques :

Les décisions n'ont pas le même impact (ni le même cout ni le même risque) et en général il faut prendre en compte cela pendant la prise de décisions :

- ✓Prêter à un client à haut risque comparativement à ne pas prêter à un client à faible risque ?
- ✓Diagnostic médical : impacts possibles de la non-détection d'une maladie grave !
- ✓Détection d'intrusion !

Pertes et risques :

Soient les définitions suivantes:

- *A={a1,a2, ...,ak}* l'ensemble des k actions (décisions) possibles(ai est la décision de classer dans la classe wi)
- *λ(ai,wj)* coût de la décision ai quand la forme appartient à la classe wj ((la perte provoquée par le mauvais classement).

D'où la règle de décision Bayésienne:

Etant donné un vecteur x caractérisant une forme

Pour i= 1,2, ... k

calculer le risque conditionnel (coût d'une décision ai) :

$$R(a_i / x) = \sum_{j=1}^{s} \lambda(a_i / w_J) \times P(w_j / x)$$

et choisir la décision *ai* telle que *R(αi/x)* soit minimum

Le choix des paramètres pour caractériser une forme:

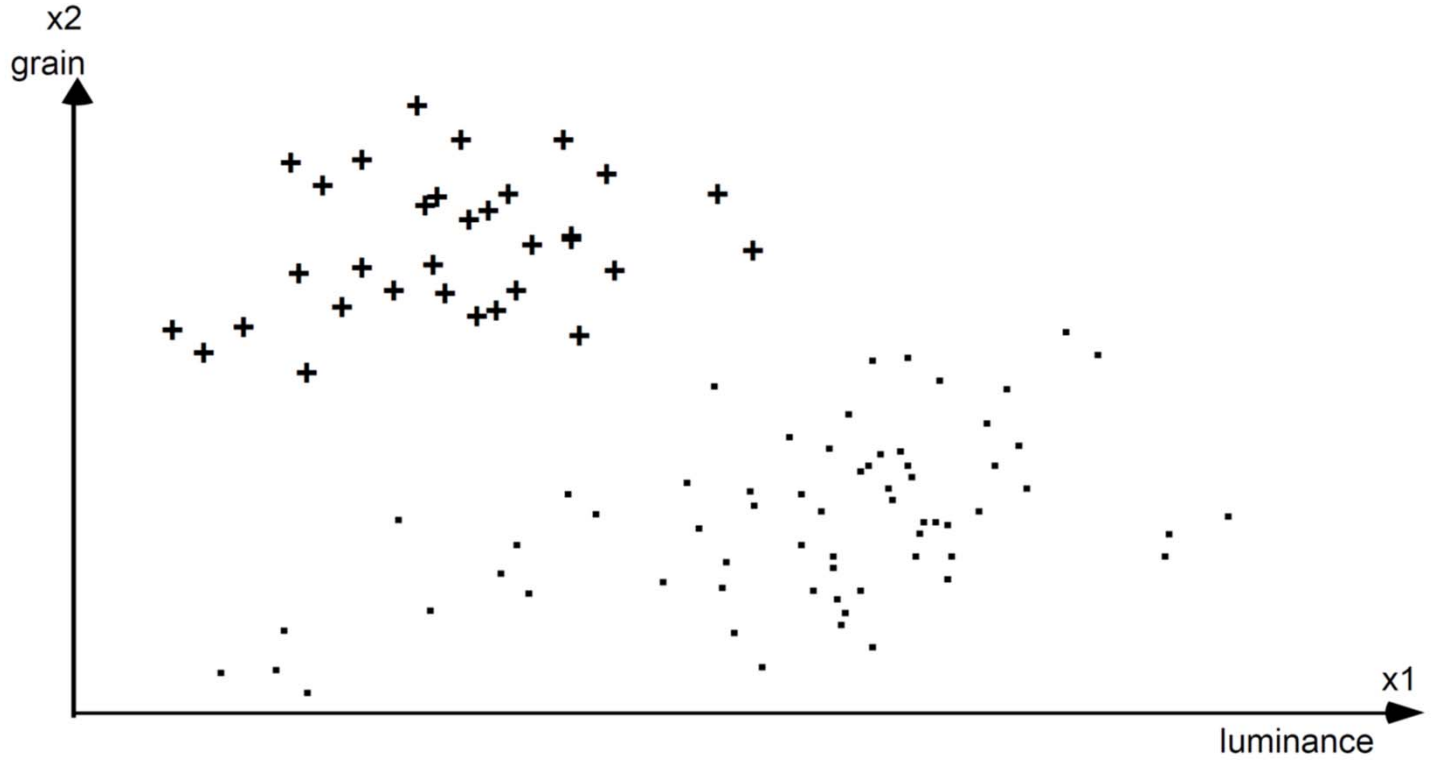
Dans l'exemple: chaque forme (une planche qui sort de la scierie) était définie par sa luminance moyenne.

Mais si nous regardons les lois de distribution des deux classes nous voyons que les deux classes ne sont pas bien séparées par ce seul paramètre puisque des échantillons eux mêmes seront mal classés par la règle de décision Bayésienne (même si on minimise la probabilité d'erreur, les erreurs seront dans ce cas de figure relativement nombreuses.

Supposons que l'on soit capable à partir de l'image fournie par la caméra de calculer le grain du bois (paramètre de texture), alors la forme sera maintenant définie par un vecteur de R2 : (x1,x2) avec x1 :luminance moyenne et x2 : grain

Le choix des paramètres pour caractériser une forme:

Si on a cette répartition pour les échantillons des classes on voit que le classifieur ne pourra être que plus performant .



Théorie de la décision bayesienne :

aborder la décision Bayésienne dans le cas le plus général

plus de 1 paramètre de mesure(la forme est définie par un vecteur de paramètres (élément de R^d).

Plus de deux classes (mais on suppose que les classes définies regroupent tous les états possibles des formes, c'est à dire que toute forme appartient obligatoirement à l'une des classes définies.

Si ce n'est pas le cas on crée une classe supplémentaire dite de « rejet » pour tomber dans cette supposition)

Loi normale et decision bayesienne

un classifieur bayésien est parfaitement défini quand on connaît pour chaque classe : **P(x/wi)** et **P(wi)**

La loi de densité de probabilité de la classe wi , p(x/wi) est alors une loi normale définie de façon analytique par quelques paramètres.

le vecteur x est en fait un nombre réel. La loi normale est dite « à une dimension »

On note la loi de densité de probabilité de la classe wi p(x/wi) sous la forme p(x) .

La loi normale notée ***N(μ,σ)*** s'écrit :

Loi normale à une dimension

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

$$avec \quad \mu = E(x) = \int_{-\infty}^{+\infty} x \times p(x) \times dx$$

$$et \quad \sigma^2 = E\left((x-\mu)^2\right) = \int_{-\infty}^{+\infty} (x-\mu)^2 \times p(x) \times x$$

Loi normale à dimension *d*

Voyons maintenant le cas où le vecteur x est dans R^d
alors on a la loi normale multi variée notée ***N(μ,Σ)***

$$p(x) = \frac{1}{(2\pi)^{d/2} \times |\Sigma|^{1/2}} \times e^{\left[-\frac{1}{2}(x-\mu)^t \times \Sigma^{-1} \times (x-\mu)\right]}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_d \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \cdot \\ \cdot \\ \mu_d \end{bmatrix}$$

Σ : matrice de covariance

$|\Sigma|$: déterminant

Avantages et limitations?

- **Avantages :**
  - le *Naive Bayes Classifier* **est très rapide pour la classification**; en effet les calculs de probabilités ne sont pas très coûteux.
  - La classification est possible même avec **un petit jeu de données**
- **Inconvénients**
  - l'algorithme *Naive Bayes Classifier* suppose **l'indépendance des variables** : C'est une **hypothèse forte** et qui est violée dans la majorité des cas réels.
  - Contre intuitivement, malgré la violation de la contrainte d'indépendance des variables, Naïve Bayes donne **de bons résultats de classification**.

Exemple :

Problème: classifier chaque personne en tant qu'individu du sexe masculin ou féminin, selon les caractéristiques mesurées. Les caractéristiques comprennent la taille, le poids, et la pointure.

- On dispose de l'ensemble de données d'entraînement suivant :

Sexe	Taille (cm)	Poids (kg)	Pointure (cm)
masculin	182	81.6	30
masculin	180	86.2	28
masculin	170	77.1	30
masculin	180	74.8	25
féminin	152	45.4	15
féminin	168	68.0	20
féminin	165	59.0	18
féminin	175	68.0	23

Exemple : suite

Sexe	Espérance (taille)	Variance (taille)	Espérance (poids)	Variance (poids)	Espérance (pointure)	Variance (pointure)
masculin	178	2.9333×10^{1}	79.92	2.5476×10^{1}	28.25	5.5833×10^{0}
féminin	165	9.2666×10^{1}	60.1	1.1404×10^{2}	19.00	1.1333×10^{1}

Nous voulons classifier l'échantillon suivant en tant que masculin ou féminin :

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

Exemple : suite

- Post (masculin) = P(masculin)*P(taille|masculin)*P(poids|masculin)*P(pointure|masculin) / évidence
- Post (féminin) = P(féminin)*P(taille|féminin)*P(poids|féminin)*P(pointure|féminin) / évidence

Le terme *évidence* (également appelé *constante de normalisation*) peut être calculé car la somme des *post* vaut 1.

évidence =
P(masculin)*P(taille|masculin)*P(poids|masculin)*P(pointure|masculin) +
P(féminin)*P(taille|féminin)*P(poids|féminin)*P(pointure|féminin)

Toutefois, on peut ignorer ce terme puisqu'il s'agit d'une constante positive (les lois normales sont toujours positives).

Exemple : suite

P(masculin) = 0.5	**P(féminin) = 0.5**
P(taille|masculin) = 4.8102e-02	**P(taille|féminin) = 7.2146e-3**
P(poids|masculin) = 1.4646e-05	**P(poids|féminin) = 3.7160e-2**
P(pointure|masculin) = 3.8052e-4	**P(pointure|féminin) = 1.1338e-1**
Postérieure (numérateur) (masculin) = 1.3404e-10	**Postérieure (numérateur) (féminin) = 1.5200e-05**

Décision : Comme la prob post *féminin* est supérieure à la prob post *masculin*, l'échantillon est plus probablement de sexe féminin.