Systèmes Intelligents : Raisonnement et Reconnaissance

James L. Crowley

Deuxième Année ENSIMAG Deuxième Semestre 2008/2009

Séance 8 8 avr 2009

Reconnaissance Bayesienne

Sources Bibliographiques :

"Neural Networks for Pattern Recognition", C. M. Bishop, Oxford Univ. Press, 1995.

"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notations

x	Une variable
X	Une valeur aléatoire (non-prévisible).
N	Le nombre de valeurs possible pour x ou X
$\vec{x}$	Un vecteur de D variables
$\vec{X}$	Un vecteur aléatoire (non-prévisible).
D	Nombre de dimensions de $\vec{x}$ ou $\vec{X}$
E	Une événement.
A,B	des classes d'événements.
T_k	La classe k
k	Indice d'une classe
K	Nombre de classes
ω_k	L'affirmation que Evennement E T_k
M_k	Nombre d'exemples de la classe k.
M	Nombre totale d'exemples de toutes les classes $M = \sum_{k=1}^{K} M_k$
$h(x)$	Histogrammes des valeurs (x est entieres avec range limité)
$h_k(x)$	Histogramme des valeurs pour la class k. $h(x) = \sum_{k=1}^{K} h_k(x)$
Q	Nombre de Cellules dans h(x). $Q = N^D$
$p(\omega_k) = p(E \quad T_k)$	Probabilité que E est un membre de la classe k.
$\vec{Y}$	La valeur d'une observation (un vecteur aléatoire).
$P(\vec{X})$	Densité de Probabilité pour $\vec{X}$
$p(\vec{X} = \vec{x})$	Probabilité q'un vecteur $\vec{X}$ prendre la valeur $\vec{x}$
$P(\vec{X} \mid \omega_k)$	Densité de Probabilité pour $\vec{X}$ etant donné que ω_k

$$P(\vec{X}) = \sum_{k=1}^{K} p(\vec{X} \mid \omega_k)\, p(\omega_k)$$

La classification

La classification est une capacité fondamentale de la vie. Pour la survie, il faut savoir reconnaître les amis, les ennemies et la nourriture.

Reconnaissance : Le fait de reconnaître, d'identifier un objet, un être comme tel.

Identifier : Reconnaître un individu

Classer : Reconnaître un membre d'une catégorie, ou d'une classe.

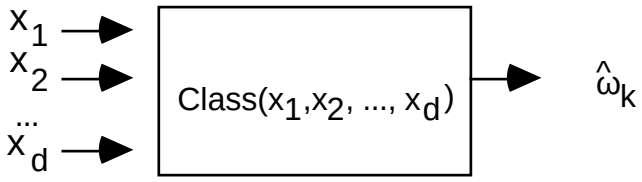
Un ensemble est défini par un test d'appartenance.

La classification est un processus d'association d'un événement à une classe. L'événement est décrit par un vecteur des caractéristiques, produit par une observation. L'affectation de l'événement à une classe est fait par un test, calculer sur le vecteur de caractéristiques.

Caractéristiques : (En anglais : Features) Signes ou ensembles de signes distinctifs.
Une ensemble de propriétés. $\{ x_1, x_2 \ldots x_D \}$.

En notation vectorielle : $\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_D \end{pmatrix}$

Pour un vecteur de caractéristique, $\vec{X}$, un processus de classification propose une estimation de la classe, $\hat{\omega}_k$



Les techniques de reconnaissance de formes, statistiques fournissent une méthode pour induire des tests d'appartenance à partir d'un ensemble d'échantillons.

Les classes peuvent êtres définis par
extension : une liste complète des membres
intention : une conjonction de prédicats.

par extension : Une comparaison d'une observation avec des membres connus de la classe (des prototypes).

Dans ce cas, la classification peut être fait par comparaison avec les membre de la classe. Soit K classe k [1, K], et M_k exemplaires pour chaque classe k : $\vec{X}_m$
Soit l'observation $\vec{Y}$:

$$k = \arg\max_k \{Sim(\vec{Y}, \vec{X}_m^{\vec{k}})\} \qquad \text{pour tout k, m} \quad \text{ou bien}$$

$$k = \arg\min_k \{ \|\vec{Y} - \vec{X}_m^k\| \} \qquad \text{pour tout k, m}$$

par intention : Conjonction de prédicats définis sur les propriétés observées. Ceci correspond (grosso modo) à des méthode dite "discriminative" de reconnaissance.

Génératives : Les techniques fondaient sur un modèle.
Discriminatives : Les techniques fondaient sur des tests quelconques.

La teste d'appartenance est une forme de partition de l'éspace de caracteristiques.

Partition de l'Espace des Caractéristiques

La classification se résume à une division de l'espace de caractéristique en partition disjoint. Cette division peut-être fait par estimation de fonctions paramétrique ou par une liste exhaustives des frontières.

Les caractéristiques définit une espace.
La classification se résume à une division de cet espace en partition disjoint de région selon la probabilité d'appartenance dans les classes.

Etant une observation, $\vec{Y}$, le critère de particition est la probabilité.

$$k = \arg\max_k \{ p(\omega_k \mid \vec{Y}) \}$$

La Probabilité d'un Evénement.

La sémantique (ou "sens") de la probabilité d'un événement peut être fourni par sa fréquence d'occurrence ou par un système d'axiomes. L'approche fréquentielle a l'avantage d'être facile à comprendre. Par contre, elle peut entraîner les difficultés dans l'analyse. La définition axiomatique favorise les analyses mathématiques.

Dans le deux cas, la probabilité est une fonction numérique, Pr() [0, 1].
Le domaine de la fonction Pr() est les evennements, E.

Définition Fréquentielle.

Une définition "Fréquentielle" de la probabilité sera suffisante pour la plupart des techniques vues dans ce cours.

Soit M observations des événement aléatoire dont M_k appartiennent à la classe A_k.
La Probabilité qu'un événement E est issue de la classe A_k est

$$p(E \quad A_k) \equiv \frac{M_k}{M}$$

La validité (ou précision) de l'approximation dépend du nombre d'échantillons M.
Si l'echantillonage est "representative", nous pouvons utiliser cette probabilité pour les evennements future.

Définition Axiomatique.

Une définition axiomatique permet d'appliquer certaines techniques d'analyse de systèmes probabilistes. Trois postulats sont suffisants :

Postulat 1 : $A_k \quad S : p(E \quad A_k) \geq 0$
Postulat 2 : $p(E \quad S) = 1$
Postulat 3 :
$A_i, A_j \quad S$ tel que $A_i \cap A_j = \quad : p(E \quad A_i \quad A_j) = p(E \quad A_i) + p(E \quad A_j)$

La probabilité de la valeur d'une variable aléatoire

Pour x entier, tel que x $\in$ $[x_{min}, x_{max}]$, on peut traiter chacun des valeurs possibles comme une classe d'événement.

Si les valeurs de x sont entières, tel que x $\in$ $[x_{min}, x_{max}]$
on peut estimer la probabilité a partir de M observations de la valeur, $\{X_m\}$.

Pour estimer la probabilité d'une valeur on peut compter le nombre d'observation de chaque valeur, x, dans une table, h(x).

L'existence des ordinateurs avec des centaines de megabytes rendre des tables de fréquence très pratique pour la mise en œuvre en temps réel des algorithmes de reconnaissance. Dans certains domaines, comme l'analyse d'images, par abus de langage, un tel table s'appelle une histogramme. Proprement dit, l'histogramme est une représentation graphique de h(x)

Ainsi la probabilité d'une valeur de X $\in$ $[X_{min}, X_{max}]$ est la fréquence d'occurence de la valeur. Avec M observations de la valeur, X, on peut faire une table, h(x), de fréquence pour chacun des valeurs possibles. On observe M exemples de X, $\{X_m\}$.

Pour chaque observation on ajoute "1" à son entré dans la table.

$$m=1, M : h(X_m) := h(X_m) + 1; \; M := M+1;$$

h(x) est une table de fréquence pour chaque x $\in$ $[x_{min}, x_{max}]$.

Ainsi, on peut définir la probabilité d'une valeur x par sa fréquence :

$$p(X_m=x) \equiv \lim_{M \to \infty} \{\frac{1}{M} h(x)\}$$

Quand M est fini, on peut faire appel à l'approximation.

$$P(X=x) \approx \frac{1}{M} h(x)$$

La validité de l'approximation depend du nombre de valeurs possible et de M.
En règle générale, on dit qu'il faut 10 exemples par cellule de l'histogramme.

La précision de cette estimation depend de la vrai densité p(X).

La pire de cas et une densité uniforme. Dans ce cas, on peut demontrer que l'ecart type moyenne de l'erreur est en proportion avec la ration de le nombre de cellule de h(x), N, sur le nombre d'echantillions, M.

$$\sigma_{MSE} \approx \frac{N}{M}$$

Que faire si la masse d'exemple est insuffisante : M << N?
Que faire si x n'est pas entier ? Il faut une fonction paramétrique pour p(X).

Densité de Probabilité.

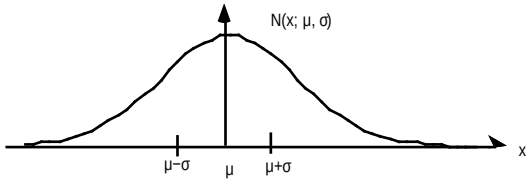
Une fonction de densité de probabilité P(X), est une fonction tel que

P(X) est Real et positive pour tout X.
X est réal entre $[-\infty, \infty]$
tel que

$$\int_{-\infty}^{\infty} P(x)\, dx = 1$$

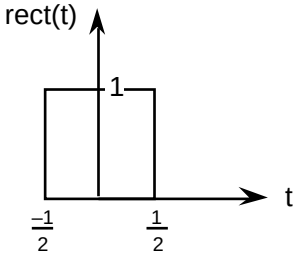
exemples :

Loi Normale $P(X) = \mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$



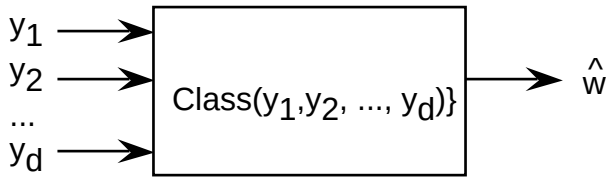
Mélange de Normales $P(X) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(x; \mu_n, \sigma_n)$

rect : P(X) = rect (X).



La classification est un processus d'association d'une observation à une classe par un teste d'appartenance.

$\hat{\omega}_k$?= Class{ $\vec{Y}$ }



Pour un vecteur de caractéristique il sort une estimation de la classe, $\hat{\omega}$

Les techniques de reconnaissance de formes statistiques fournissent une méthode pour induire des tests d'appartenance à partir d'un ensemble d'échantillons.

La classification se résume à une division de l'espace de caractèristique en partition disjoint. Cette division peut-être fait par estimation de fonctions paramétrique ou par une liste exhaustives des frontières.

Le critère est la probabilité. $\hat{\omega}_k = \underset{\omega_k}{\text{arg-max}}\{ p(\omega_k \mid \vec{X}) \}$

Cette probabilité est fournie par la règle de Bayes.

$$p(\omega_k \mid \vec{X}) = \frac{p(\vec{X} \mid \omega_k)\, p(\omega_k)}{p(\vec{X})}$$

La Règle de Bayes

Soit un evennement "E". Soit deux tribus d'evennements A et B tel que certains evennements sont common à A et à B.
E peut appartenir à $A \cap B$ ou a $\neg A \cap B$ ou a $A \cap \neg B$ ou a $\neg A \cap \neg B$

Soit deux propositions p et q.

p ≡ E A et q ≡ E B
donc P(p) ≡ Pr{E A} et P(q) ≡ Pr{E B}.

Par axiome 2 de la définition des systèmes de probabilités :

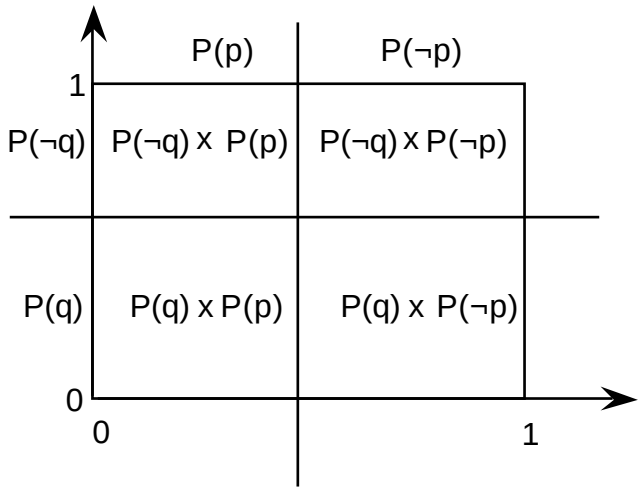
$P(q) + P(\neg q) = 1.$

P(p q) est la probabilité "conjointe" de p et q.

Si p et q sont independentes

$P(p \quad q) = P(p) \cdot P(q),$
$P(p \quad q) = P(p) + P(q).$

On peut voir ça d'une manière graphique :



$P(p \quad q) + P(p \quad \neg q) + P(\neg p \quad q) + P(\neg p \quad \neg q) = 1$

Dans ce cas, les probabilités marginales sont

$$P(p) = P(p \quad q) + P(p \quad \neg q)$$
$$P(q) = P(p \quad q) + P(\neg p \quad q)$$

La probabilité conditionnelle de q étant donnée p s'écrit $P(q \mid p)$

$$P(q \mid p) = \frac{P(p \quad q)}{P(p)} = \frac{P(p \quad q)}{P(p \quad q) + P(p \quad \neg q)}$$

de la même manière :

$$P(p \mid q) = \frac{P(p \quad q)}{P(q)} = \frac{P(p \quad q)}{P(p \quad q) + P(\neg p \quad q)}$$

Par algèbre on déduire :

$$P(q \mid p)\, P(p) = P(p \quad q) = P(p \mid q)\, P(q)$$

d'ou

$$P(q \mid p)\, P(p) = P(p \mid q)\, P(q)$$

Ceci est une forme de règle de Bayes. On peut écrire :

$$P(q \mid p) = \frac{P(p \mid q)\, P(q)}{P(p)}$$

$P(q \mid p)$ est la probabilité "conditionnelle" ou "postérieur"

La règle de Bayes avec une ratio d'histogrammes.

On peut utiliser la règle de Bayes pour calculer la probabilité d'apartenance d'un classe. Soit un vecteur de caractéristiques, X discrètes tel que x $[X_{min}, X_{max}]$, et une ensemble de classe ω_k.

$$p(\omega_k \mid X=x) = \frac{p(X=x \mid \omega_k)}{P(X=x)} p(\omega_k)$$

probabilité de la classe ω_k: $p(\omega_k) \approx \frac{M_k}{M}$

probabilité conditionnelle de X): $p(X=x \mid \omega_k) \approx \frac{1}{M_k} \; h_k(x)$

Probabilité à priori de X : $p(X=x) \approx \frac{1}{M} \; h(x)$

ce qui donne :

$$p(\omega_k \mid X=x) = \frac{p(X=x \mid \omega_k)}{p(X=x)} p(\omega_k) \approx \frac{\frac{1}{M_k} h_k(x)}{\frac{1}{M} h(x)} \frac{M_k}{M} = \frac{h_k(x)}{h(x)}$$

Cette technique peux également marcher pour les vecteurs de caractéristiques.

La histogramme est une table a D dimensions : $h(\vec{x})$

probabilité conditionnelle de $\vec{X}$: $p(\vec{X}=\vec{x} \mid \omega_k) \approx \frac{1}{M_k} \; h_k(\vec{x})$

Probabilité à priori de $\vec{X}$: $p(\vec{X}=\vec{x}) \approx \frac{1}{M} \; h(\vec{x})$

ce qui donne :

$$p(\omega_k \mid \vec{X}=\vec{x}) = \frac{p(\vec{X}=\vec{x} \mid \omega_k)\, p(\omega_k)}{p(\vec{x})} \approx \frac{\frac{M_k}{M} \frac{1}{M_k} h_k(\vec{x})}{\frac{1}{M} h(\vec{x})} = \frac{h_k(\vec{x})}{h(\vec{x})}$$

Cette technique s'avère très utile dans les cas où il y a suffisemment d'échantillons pour faire un histogramme valable. Par exemple quand on traite des images ou les signaux.

Exemple : Les statistiques de couleurs de la peau

Une image est une table de pixels.
Chaque pixel est une observation d'une scène, et donc, une variable aléatoire.
Il y a beaucoup des pixels dans les images (512 x 512 = 2^{18} = 256 K pixels)

Les pixels d'une image couleur sont representé par 3 octets R, G et B avec (8 bits par octets). Dans ce cas, chaque pixel est une vecteur aléatoire.

$$\vec{X} = (R, G, B)^T$$

ou R, G, et B sont issue du [0, 255].

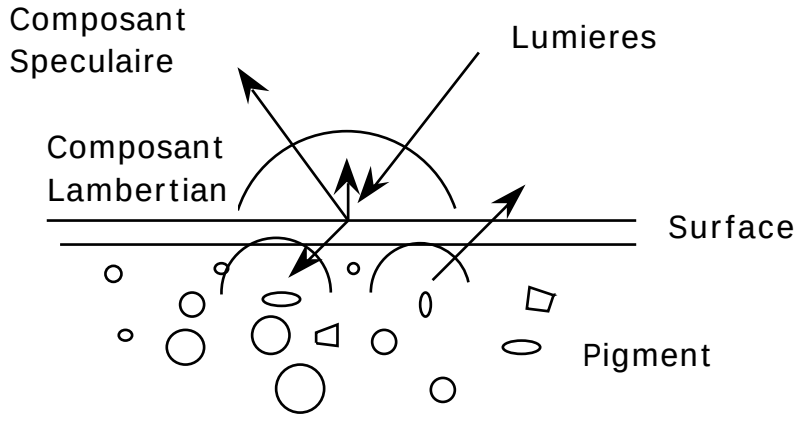
Pour un vecteur de caractéristique, on peut calculer une table à 3 dimensions.

Pour un image couleur, composé de (R, G, B), avec 8 bits par pixel, $h(\vec{X})$ contient $256^3 = 2^{24}$ valeurs. Mais chaque image contient $512^2 = 2^{18}$ pixels.
Si on suppose qu'il faut 10 exemples par cellulle, Il faut 10 x 2^6 images = 640 images pour une estimation valable de $p(\vec{X}) = \frac{1}{M} h(\vec{X})$.

A 25 images per second, ceci est 25.6 seconds de vidéo.

Si on souhaite, on peut reduire le nombe de dimensions, en normalisant la luminance.

On peut transformer le vecteur $(R, G, B)^T$ en luminance et chrominance.
La luminance, ou intensité, L, est en proportion de cos(i) où i est l'angle entre la source et la normale de la surface. La chrominance, C_1, C_2 est une signature pour la reconnaissance.



La composant "luminant" est déterminé par l'orientation de la surface.

La composant "chrominant" est déterminé par la composition de la spectre de la source et le spectre d'absorbation des pigments de la surfaces. Si la spectre de la source est constante, la chrominance indique l'identité de l'objet

Par exemple :

$$L = R+G+B \quad C_1 = \frac{R}{R+G+B} \qquad C_2 = \frac{G}{R+G+B}$$

R, G, B sont les entiers. Donc, C_1, C_2 sont issu d'une ensemble finit de valeurs dans l'intervalle [0, 1]. On peut transformer C_1, C_2 en entier entre [0, N-1], par

$$c_1 = \text{Round}\left(N \cdot \frac{R}{R+G+B} \right). \qquad c_2 = \text{Round}\left(N \cdot \frac{G}{R+G+B} \right).$$

Donc pour chaque pixel (i,j), $\vec{Y} = (C_1, C_2)$

On aura N^2 cellulles de chrominances dans l'histogramme.
Par exemple, pour N=32, on a $32^2 = 1024$ cellules à remplir est
il nous faut que M = 10 K pixels d'exemples. (Une image = 256 K pixels).

Dans ce cas, pour M observations $p(\vec{X}=\vec{x}) \approx \frac{1}{M} \; h(\vec{x})$

Un histogramme de couleurs, $h(\vec{X})$, de les M pixels dans une l'image donne une approximation de la probabilité de chaque couleur dans l'image.

$$p(\vec{X}) \equiv p(\vec{X}=\vec{x}) \approx \frac{1}{M} \; h(\vec{x})$$

Un histogramme des de couleurs d'un entité, A , $h_A(\vec{X})$, de les M_A pixels dans une région d'une image de l'objet, w(i, j), donne une approximation de la probabilité de chaque couleur de l'objet.

$$p(\vec{X} \mid A) \equiv p(\vec{X}= \vec{x} \mid A) \approx \frac{1}{M_A} \; h_A(\vec{x})$$

Détection par ratio d'histogramme

L histogramme permet d'utiliser la règle de Bayes afin de calculer
la probabilité qu'un pixel corresponde à un objet.

Pour chaque pixel de chrominance $\vec{X}(i, j)$: $p(\text{objet} \mid \vec{x}) = p(\vec{x} \mid \text{objet}) \frac{p(\text{objet})}{p(\vec{x})}$

Soit N images de M pixels. Ceci fait M Pixels.
Soit $h(c_1, c_2)$, l'histogramme de tous les M pixels.
Soit $h_A(c_1, c_2)$, l'histogramme des M_A pixels de l'objet "A".

$$p(c(i,j) = A) = \frac{M_A}{M} \qquad p(\vec{x}) \approx \frac{1}{M} h(\vec{x})$$

$$p(\vec{x} \mid A) \approx \frac{1}{M_A} h_A(\vec{x})$$

Donc $p(A \mid \vec{x}) = p(\vec{x} \mid A) \frac{p(A)}{p(\vec{x})} = \frac{1}{M_A} h_A(\vec{x}) \frac{\frac{M_A}{M}}{\frac{1}{M} h(\vec{x})}$

$$p(A \mid \vec{Y}) = \frac{h_A(\vec{x})}{h(\vec{x})}$$

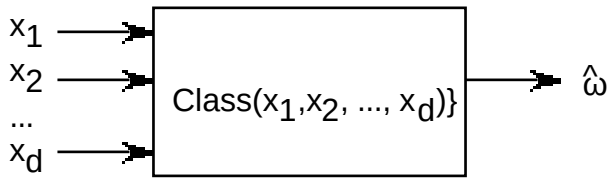
Il faut assurer que $h(\vec{x}) \neq 0$!! Pour cela il faut que $w_A(i,j)$ $w(i,j)$.

Ainsi, on peut créer une image de probabilité de l'objet.
En peut ensuite déterminer un seuil pour l'image de probabilité.

La Classification

Soit les événements E décrit par une vecteur de caractéristiques $\vec{X}$: $(E,\vec{X})$.
Soit K classes d'événements $\{T_k\} = \{T_1, T_2, ..., T_K\}$

La classification est un processus d'estimation de l'appartenance d'un événement à une des classes T_k fondée sur les caractéristiques de l'événement, $\vec{X}$.



$\hat{\omega}_k$ = Decider(E T_k)

$\hat{\omega}_k$ est la proposition que (E T_k).
La fonction de classification est composée de deux parties d() et $g_k()$:

$\hat{\omega}_k = d(g(\vec{X}))$.

$g(\vec{X})$: Une fonction de discrimination : $R^D \rightarrow R^K$
d() : Une fonction de décision : $R^K \rightarrow \{\omega_K\}$

La Classification Bayesienne

La technique Bayesienne de Classification repose sur une fonction de vérité probabiliste et le règle de Bayes.

Dans un système de vérité probabilisté, la valeur de vérité de la proposition une probabilité :

$$p(\omega_k) = p(E \quad T_k)$$

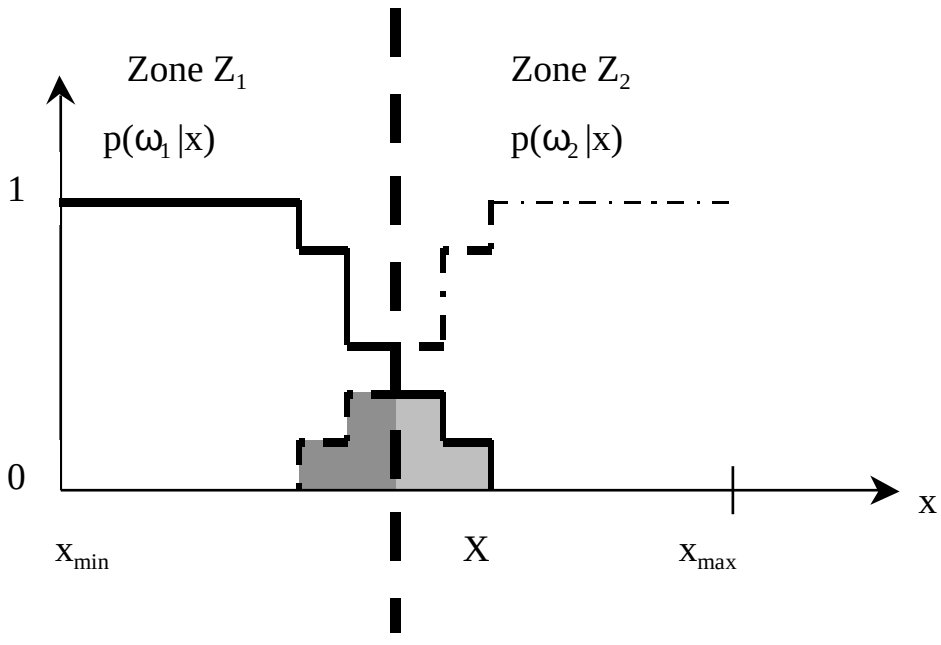
Le critère de décision est de minimiser le nombre d'erreur. Dans un système probabiliste, ca revient de minimiser la probabilité d'erreur. Ceci est équivalent à choisir la classe le plus probable.

$$\hat{\omega}_k = \text{Decider}(E \quad T_k). = \arg\max_{\omega_k} \{p(\omega_k \mid \vec{X})\}$$

Pour estimer la probabilité nous utilisons les caractéristiques, $\vec{X}$, de l'événement.

Conisdère le cas D = 1 et K = 2. Dans ce cas, le domaine d'X est un axe.
La classification est équivalente à une decoupage du domaine d'X en deux zones : Z_1 et Z_2.

$\hat{\omega}_1$ si X Z_1 et $\hat{\omega}_2$ si X Z_2



La probabilité d'erreur est la somme des probabilités de $p(\omega_2)$ en Z1 et

la somme de probabilité de $p(\omega_1 | X)$ en zone 2.

$$p(\text{erreur}) = \sum_{Z_1} p(\omega_2 | X) + \sum_{Z_2} p(\omega_1 | X)$$

La minimum est atteint quand :

Donc $d(g_k(X)) = \arg\max_{\omega_k}\{p(\omega_k | X)\}$

Dans ce cas, nous avons utilisé $\arg\max_{\omega_k}\{p(\omega_k | X)\}$ en tant que fonction de décision et $g_k(X) = p(\omega_k | X)$ comme la fonction de discrimination