Vision par Ordinateur

James L. Crowley

DEA IVR Premier Bimestre 2005/2006

Séance 5 9 novembre 2005

Reconnaissance de formes

Plan de la Séance :

Notations

$\vec{x}$	Un vecteur
$\vec{X}$	Un vecteur aléatoire (non-prévisible).
D	Nombre de dimensions de $\vec{X}$
w_k	La classe k
k	Indice d'une classe
K	Nombre de classes
M_k	Nombre d'exemples de la classe k.
M	Nombre totale d'exemples de toutes les classes
$p(w_k)$	Probabilité a priori de rencontrer un membre de la classe k.
$\vec{Y}$	Une observation (un vecteur aléatoire).
$P(\vec{Y})$	Probabilité d'une observation $\vec{Y}$

La Classification des Formes

La classification est une capacité fondamentale de l'intelligence.

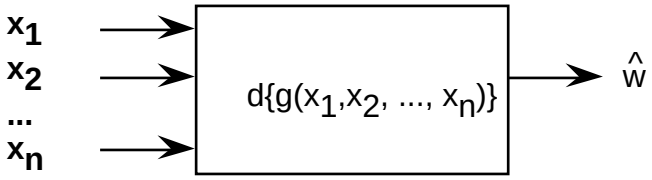
Reconnaissance : Le fait de reconnaître, d'identifier un objet, un être comme tel.

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'une observation à une classe par un teste d'appartenance.



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

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 d'appartenance.
Cette probabilité est fournie par la règle de Bayes.

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

Les Formes

Forme n. f. : A. Apparence, aspect visible. 1) ... 2) apparence extérieure donnant à un objet ou à un être sa spécificité.

Les méthodes statistique de la reconnaissance de forme traite les observations sous forme de vecteur de caractéristiques.

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

En notation vectorielle :

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_D \end{pmatrix}$$

La formation des vrais objets physiques est sujette aux influences aléatoires. Pour les objets d'une classe, w_k, les propriétés des objets individuels sont, les valeurs aléatoires. On peut resume ceci par une somme d'une forme "intrinsèque" $\vec{x}$ plus ces influences aléatoires individuelles, $\vec{B}_i$.

$$\vec{X} = \vec{x} + \vec{B}_i$$

Les techniques probabiliste de RF suppose un bruit additif.

En notation vectorielle :

$$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \ldots \\ X_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \ldots \\ x_n \end{pmatrix} + \begin{pmatrix} B_1 \\ B_2 \\ \ldots \\ B_n \end{pmatrix}$$

Les Observations

Une observation : une constatation attentive des phénomènes.

Pour des machines, des observations sont fournies par les capteurs.

Ceci donne une observation (un phénomène) sous forme d'une ensemble de caractéristiques : $\{ Y_1, Y_2 \ldots Y_D\}$.

$$\vec{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix}$$

Les observations sont corrompues par un bruit, $\vec{B}_o$.

$$\vec{Y} = \vec{y} + \vec{B}_o$$

Le bruit est, par définition, imprévisible. Il est aléatoire.
Donc les caractéristiques observées sont des vecteurs aléatoires.
La corruption des observations par un bruit aléatoire est fondamentale aux capteurs physiques.

Pour chaque classe w_k, la probability d'observé $\vec{Y}$ est fournie par la règle de Bayes.

$$p(w_k \mid \vec{Y}) = \frac{p(\vec{Y} \mid w_k)\, p(w_k)}{p(\vec{Y})}$$

Si $\vec{Y} = \vec{X} + \vec{B}_o$ est issu de la classe w_k ayant characteristique $\vec{X} = \vec{x} + \vec{B}_i$

Par exemple :

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

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

$$C_1 = \text{Round}\left(N \cdot \frac{R}{R+G+B} \right). \qquad C_2 = \text{Round}\left(N \cdot \frac{G}{R+G+B} \right).$$

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

La probabilité à posteriori peut être calculé par la règle de Bayes.

$$p(w_k \mid \vec{X}) = \frac{p(\vec{X} \mid w_k)\; p(w_k)}{p(\vec{X})}$$

Dans le cas des valeurs de $\vec{X}$ discrètes tel que x $[X_{min}, X_{max}]$, on a

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

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

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

ce qui donne :

$$p(w_k \mid \vec{X}) = \frac{p(\vec{X} \mid w_k)\; p(w_k)}{p(\vec{X})} \approx \frac{\frac{M_k}{M} \; \frac{1}{M_k} \; h_k(\vec{X})}{\frac{1}{M} \; h(\vec{X})} = \frac{h_k(\vec{X})}{h(\vec{X})}$$

Que faire si la masse d'exemple est insuffisante : $M < 10\,(X_{max} - X_{min})$?
Que faire si x n'est pas entier ? Il faut une fonction paramétrique pour p(X).

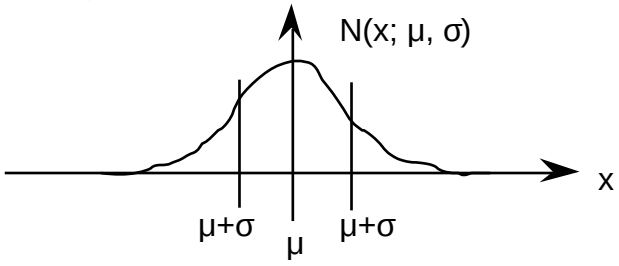
La Loi Normale

Quand les variables aléatoires sont issues d'une séquence d'événements aléatoires, leur densité de probabilité prend la forme de la loi normale, $\mathcal{N}(\mu, \sigma)$. Ceci est démontré par le théorème de la limite centrale. Il est un cas fréquent en nature.

Les paramètres de $\mathcal{N}(\mu, \sigma)$ sont les premiers et deuxième moments des exemples. Donc, on peut les estimer pour n'importe quel nombre d'exemples. On peut même estimer les moments quand il n'existe pas les bornes (X_{max}–X_{min}) ou quand X est une variable continue.

Dans ce cas, p( ) est une "densité" et il faut une fonction paramétrique pour p(). Dans la plupart des cas, on peut utiliser $\mathcal{N}(\mu, \sigma)$ comme une fonction de densité pour p(x).

$$p(x) \approx \mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \; e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Le base "e" est : e = 2.718281828....Il s'agit du fonction tel que $\int e^x \, dx = e^x$

Le terme $\frac{1}{\sqrt{2\pi}\sigma}$ sert à normaliser la fonction en sorte que sa surface est 1.

$$\int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \sqrt{2\pi}\,\sigma.$$

Le terme $d^2(x) = \frac{(x-\mu)^2}{\sigma^2}$ est la distance x et μ normalisée par la variance.

La différence $(x - \mu)^2$ est la "distance" entre une caractéristique et la caractéristique "nominale" d'une classe. La variance, σ^2, sert à "normaliser" cette distance.

La différence normalisée par la variance est connue sous le nom de "Distance de Mahalanobis". La Distance de Mahalanobis est un test naturel de similarité

Estimations des moments d'une densité

Le premier moment : La Moyenne

Soit M observations d'un variable aléatoire, $\{ X_1, X_2, ..., X_3\} = \{X_m\}$
La moyenne est l'espérance de $\{X_m\}$. .

$$\mu \equiv E\{x\} = \frac{1}{M} \sum_{m=1}^{M} X_m$$

Il s'agit d'une somme sur M (le nombre exemples). Cette somme existe, même quand X_{min} et X_{max} n'existent pas et quand X est réelle.

On note que dans le cas X est un nombre entier, on peut aussi estimer la moyenne par la table de fréquence. La masse d'un histogramme, h(x) est le nombre d'échantillons qui composent l'histogramme, M.

$$M = \sum_{x=x_{min}}^{x_{max}} h(x)$$

Pour X entier, tel que $X \quad [x_{min}, x_{max}]$ on peut démontrer que

$$\mu \equiv E\{X\} = \frac{1}{M} \sum_{x=x_{min}}^{x_{max}} h(x) \approx \sum_{x=x_{min}}^{x_{max}} p(x) \cdot x$$

donc : $\mu \equiv E\{X\} = \frac{1}{M} \sum_{m=1}^{M} X_m = \frac{1}{M} \sum_{x=x_{min}}^{x_{max}} h(x)$

Pour X continue :

$$\mu \equiv E\{X\} = \int p(x) \cdot x \, dx$$

Le deuxième moment (La variance)

La variance σ^2 est le deuxième moment de la densité de probabilité.
Pour un ensemble de M observations $\{x_m\}$

$$\sigma^2 \equiv E\{(X_m-\mu)^2\} = \frac{1}{M}\sum_{m=1}^{M}(X_m-\mu)^2$$

Mais l'usage de μ estimé avec le même ensemble, introduit un biais dans σ^2.
Pour l'éviter, on peut utiliser une estimation sans biais.

$$\sigma^2 = \frac{1}{M-1}\sum_{m=1}^{M}(X_m-\mu)^2$$

Lequel est correct ? (les deux !)

Pour X entier, tel que X $[X_{min}, X_{max}]$ on peut démontrer que

$$\sigma^2 = E\{(X_m-\mu)^2\} = \frac{1}{M}\sum_{x=x_{min}}^{Xmax} h(x)(x-\mu)^2$$

Ceci est vrai par ce que la table h(x) est faite de $\{X_m\}$.

Donc :

$$\sigma^2 = \frac{1}{M}\sum_{m=1}^{M}(X_m-\mu)^2 = \frac{1}{M}\sum_{x=x_{min}}^{xmax} h(x)(x-\mu)^2$$

Pour X réel on a

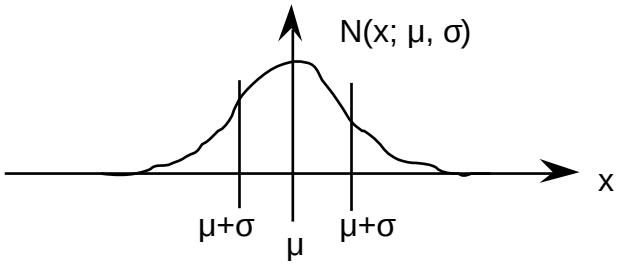
$$\sigma^2 \equiv E\{(X_m-\mu)^2\} = \int p(x)\cdot(x-\mu)^2\,dx$$

La Loi Normale pour D = 1

Avec μ et σ^2, on peut estimer la densité p(x) par $\mathcal{N}(x; \mu, \sigma)$

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

$\mathcal{N}(x; \mu, \sigma)$ a la forme :



La moyenne est le premier moment de la densité p(x).

$$\mu \equiv E\{X\} = \int p(x) \cdot x \, dx$$

La variance σ^2 est le deuxième moment de p(x).

$$\sigma^2 \equiv E\{(X-\mu)^2\} = \int p(x) \cdot (x-\mu)^2 \, dx$$

La Loi Normale pour D > 1

Pour un vecteur de D propriétés

$$\vec{\mu} \equiv E\{\vec{X}\} = \frac{1}{M}\sum_{m=1}^{M} \vec{X}_m = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix} = \begin{pmatrix}E\{X_1\}\\ E\{X_2\}\\ \dots\\ E\{X_D\}\end{pmatrix}$$

Pour $\vec{X}$ entier, tel que pour chaque d [1, D], x_d $[x_{dmin}, x_{dmax}]$ on peut démontrer que

$$\mu_d \equiv E\{x_d\} = \frac{1}{M}\sum_{x_1=x_{1min}}^{x_{1max}} \dots \sum_{x_D=x_{Dmin}}^{x_{Dmax}} h(\vec{x})\, x_d$$

Pour x réel, $\mu_d \equiv E\{x_d\} = \int\int \dots \int p(\vec{x}) \cdot x_d \, d\vec{X}$

Dans tous les cas :

$$\vec{\mu} = E\{\vec{\mathbf{X}}\} = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_n\end{pmatrix} = \begin{pmatrix}E\{x_1\}\\ E\{x_2\}\\ \dots\\ E\{x_n\}\end{pmatrix}$$

Pour D dimensions, la covariance entre les variables x_i et x_j est estimée à partir de M observations $\{\vec{\mathbf{X}}_m\}$

$$\sigma_{ij}^2 \equiv E\{ (X_i - E\{X_i\})(X_j - E\{X_j\})\}$$

$$= \frac{1}{M}\sum_{m=1}^{M} (X_{im} - \mu_i)(X_{jm} - \mu_j))$$

Et encore, pour éviter le bias, on peut utiliser :

$$\sigma_{ij}^2 = \frac{1}{M-1}\sum_{m=1}^{M} (X_{im} - \mu_i)(X_{jm} - \mu_j))$$

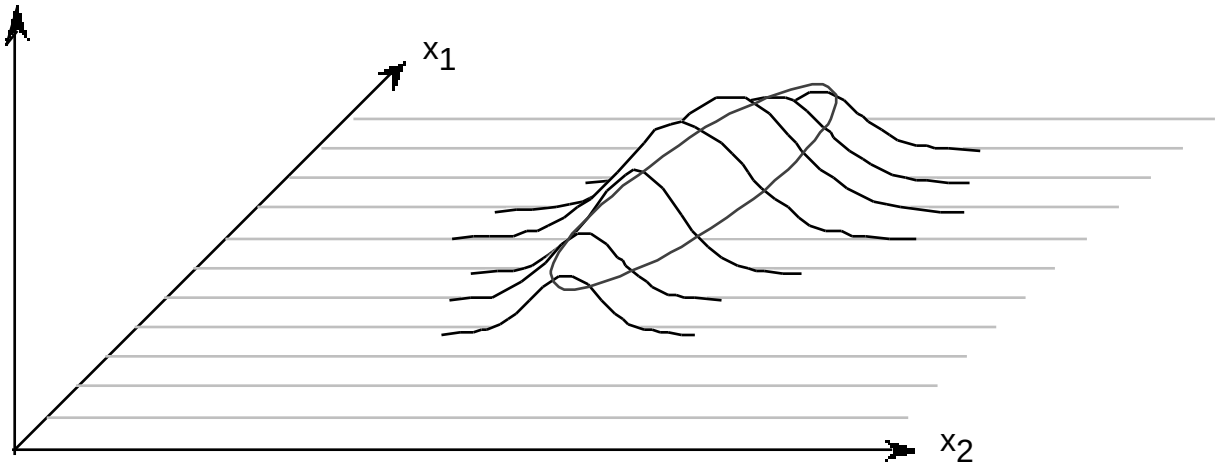
Ces coefficients composent une matrice de covariance. C

$$\mathbf{C}_X \equiv E\{[\vec{\mathbf{X}} - \vec{\mu})(\vec{\mathbf{X}} - \vec{\mu})^T\} = E([\vec{\mathbf{X}} - E\{\vec{\mathbf{X}}\})(\vec{\mathbf{X}} - E\{\vec{\mathbf{X}}\})^T\}$$

$$\mathbf{C}_x \equiv \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots & \sigma_{1D}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots & \sigma_{2D}^2 \\ \dots & \dots & & \dots \\ \sigma_{D1}^2 & \sigma_{D2}^2 & \dots & \sigma_{DD}^2 \end{pmatrix}$$

Dans le cas d'un vecteur de propriétés, $\vec{X}$, la loi normale prend la forme :

$$p(\vec{X}) = \mathcal{N}(\vec{X}; \vec{\mu}, C) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\mathbf{C})^{\frac{1}{2}}} \; e^{-\frac{1}{2}(\vec{X} - \vec{\mu})^T C_X^{-1} (\vec{X} - \vec{\mu})}$$



Le terme $(2\pi)^{\frac{D}{2}} \det(\mathbf{C})^{\frac{1}{2}}$ est un facteur de normalisation.

$$\int\int \dots \int e^{-\frac{1}{2}(\vec{X} - \vec{\mu})^T C_X^{-1} (\vec{X} - \vec{\mu})} \, dX_1 \, dX_2 \dots dX_D = (2\pi)^{\frac{D}{2}} \det(\mathbf{C})^{\frac{1}{2}}$$

La déterminante, det(C) est une opération qui donne la "énergie" de C.

Pour D=2 $\det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a \cdot b - c \cdot d$

Pour D=3

$$\det \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} = a \cdot \det \begin{pmatrix} e & f \\ h & i \end{pmatrix} + b \cdot \det \begin{pmatrix} f & d \\ i & g \end{pmatrix} + c \cdot \det \begin{pmatrix} d & e \\ g & h \end{pmatrix}$$

$$= a(ei - fh) + b(fg - id) + c(dh - eg)$$

pour D > 3 on continue récursivement.

L'exposant est une valeur positive et quadrique.
(si X est en mètre, $\frac{1}{2}(\vec{\mathbf{X}} - \vec{\mu})^{T}\mathbf{C}_X^{-1}(\vec{\mathbf{X}} - \vec{\mu})$ est en mètre$^{2.}$)

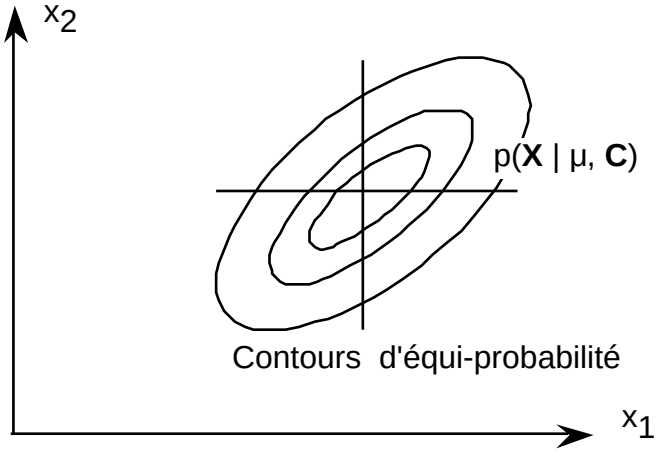
Cette valeur est connue comme la "distance de Mahalanobis".

$$d^2(\vec{X}) = \frac{1}{2}(\vec{X} - \vec{\mu})^{T}C_X^{-1}(\vec{X} - \vec{\mu})$$

Il s'agit d'une distance euclidienne, normalisé par la covariance C_X.

Cette distance est bien définie, même si les composants de $\vec{X}$ n'ont pas les mêmes unités. (Ceci est souvent le cas).

La loi Normale peut être visualisé par ses contours d'"équiprobabilité"



Ces contours sont les contours de constant $d^2(\vec{X})$

La matrice C est positive et semi-definite. Nous allons nous limiter au cas ou C est positive et definite (C.-à-d. $\det(C) = |C| > 0$

si x_i et x_j sont statistiquement indépendants, $\sigma_{ij}^2 = 0$.

si pour tout $i \neq j$, σ_{ij}^2 alors $p(\vec{X}) = \prod_{i=1}^{n} p(X_i)$

La Classification

Classification entre deux Catégories (K=2, D=1)

Soit un événement A avec propriété X.
Soit w_1 une tribu d'événements.

On note que A w_1 ou A w_1

soit la proposition p que X vaut la valeur "x".

p : X= x. on éctir : $p(X) = Pr\{X=x\}$

Soit une classe w_1. Soit A w_1 ou A w_1
soit proposition q : A w_1 donc $\neg q$: A w_1

on écrit : $p(q) = Pr\{A \; w_1\} = p(w_1)$

Bayes : $p(q \mid p)\, p(p) = p(p \mid q)\, p(q)$

$Pr(A \; w_1 \mid X=x)\, Pr\{X=x\} = Pr(X \mid A \; w_1)\, Pr(A \; w_1) =$

ou bien : $p(w_1 \mid X)\, p(X) = p(X \mid w_1)\, p(w_1)$

Ceci donne :

$$p(w_1 \mid X) = \frac{p(X \mid w_1)\, p(w_1)}{p(X)}$$

Soit une deuxième classe w_2 pour tous les événements qui n'est pas de w_1.

A $w_1 \Leftrightarrow$ A w_2

Les classes w_1 et w_2 sont mutuellement exclusifs.

Soit x. Comment décider entre w_1 et w_2 ?

Une idée simple serait de chercher à minimiser la probabilité d'erreur.

$$p(\text{Erreur} \mid X) = \begin{cases} p(w_1 \mid X) = Pr\{A \; w_1 \mid X=x\} & \text{si on decide } w_2 \\ p(w_2 \mid X) = Pr\{A \; w_2 \mid X=x\} & \text{si on decide } w_1 \end{cases}$$

Donc pour tout X : p(Erreur | X) est minimale si on décide le w_k tel que

si $p(w_1 | X) > p(w_2 | X)$ décide w_1 sinon décide w_2

Par règle de Bayes ceci est

si $p(X | w_1)\, p(w_1) > p(X | w_2)\, p(w_2)$ decide w_1 sinon décide w_2

L'idée intuitive est de voir les valeurs de la propriété x comme un espace. Pour une scalaire x (D = 1) l'espace est une droite. (Une droite est une ensemble ordonnées de points.)

La frontière $p(x | w_1)\, p(w_1) = p(x | w_2)\, p(w_2)$ partition l'espace x dans deux régions disjointes, z_1 et z_2.

La Probabilité d'erreur

Notre règle

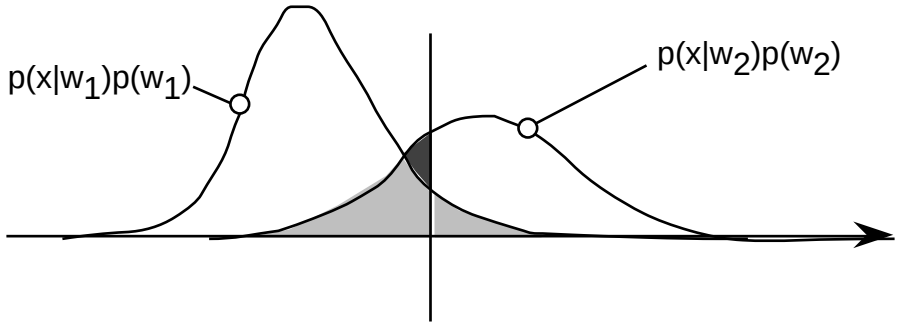
si $p(X | w_1)\, p(w_1) > p(X | w_2)\, p(w_2)$ decide w_1 sinon décide w_2

Notre surface de décision est $p(X | w_1)\, p(w_1) = p(X | w_2)\, p(w_2)$

On peut noter que

$$p(\text{Erreur} | X) = p(X \in z_2, w_1)p(w_1) + p(X \in z_1, w_2)p(w_2)$$

$$= \int_{z_2} p(x | w_1)\, p(w_1)\, dx + \int_{z_1} p(x | w_2)\, p(w_2)\, x$$



La valeur minimum de P(erreur) est atteinte pour $\frac{d\, P(\text{erreur})}{dx} = 0$

Donc quand $p(\vec{x} | w_1)\, p(w_1) - p(\vec{x} | w_2)\, p(w_2) = 0$.

Classification pour K > 2

Soit D=1. Dans le cas de K > 2 , il y a plus de possibilité d'erreurs.

Il vaut mieux maximiser $P(\text{vrai} \mid X) = p(\neg \text{Erreur} \mid X)$

Dans le cas général, on sélection la classe w_k pour laquelle

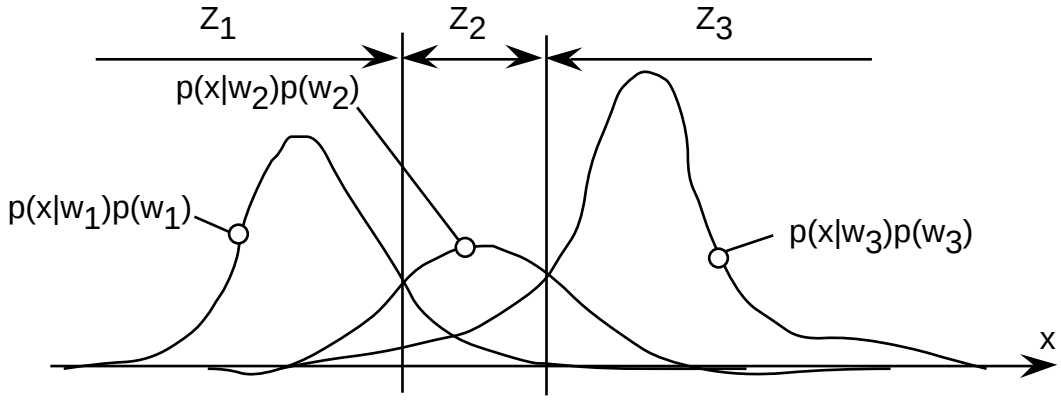
$$k = \arg\max_k \{g_k(X)\} \qquad \text{avec } g_k(X) = p(w_k \mid X)$$

par règle de Bayes :

$$\arg\max_k \{p(w_k \mid X)\} = \arg\max_k \left\{\frac{p(X \mid w_k)\, p(w_k)}{p(X)}\right\}$$

Les frontières entre régions i et j sont les valeurs pour lesquelles

$$g_i(X) = g_j(X)$$



Une fonction de discrimination partition l'espace de caractéristique en régions disjointes $Z_1, \ldots, Z_k$ pour chaque classe.

Classification :

$$w_k = \arg\max_k \{g_k(\vec{X})\}$$

Classification pour K > 2 et D > 1.

Dans le cas général, $g_k(\vec{X}) = p(w_k|\vec{X})\, p(w_k)$

Soit $g_k(\vec{X}) = p(w_k \mid \vec{X})\, p(w_k)$.

Quand l'observation est un vecteur (D> 1) on aura

si $i \neq j$ $g_i(\vec{X}) > g_j(\vec{X})$ decide w_i

Dans cette forme la classificateur est une machine qui calcule K fonctions $g_k(\vec{x})$ suivie d'une sélection du maximum.

La fonctions de discrimination est : $g_k(\vec{X}) = p(w_k \mid \vec{X}) = \dfrac{p(\vec{X} \mid w_k)\, p(w_k)}{p(\vec{X})}$

On sélection la classe w_k pour laquelle $\arg\max_k\{g_k(\vec{X})\}$

par règle de Bayes :

$$\arg\max_k\{p(w_k \mid \vec{X})\} = \arg\max_k\{\frac{p(\vec{X} \mid w_k)\, p(w_k)}{p(\vec{X})}\}$$

$= \arg\max\{ p(\vec{X} \mid w_k)\, p(w_k)\}$ parce que $p(\vec{x})$ est indépendant de w_k

$= \arg\max\{ \mathrm{Log}\{p(\vec{X} \mid w_k)\} + \mathrm{Log}\{p(w_k)\}$ parce que Log{} est une fonction monotone.

Si le bruit est d'une densité Normale :

$$p(\vec{X} \mid w_k) = \mathcal{N}(\vec{\mu}_k, C_k)$$

$$\mathrm{Log}\{p(\vec{X} \mid w_k)\} = \mathrm{Log}\{\frac{1}{(2\pi)^{\frac{D}{2}} \det(C_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1}(\vec{X}-\vec{\mu}_k)}\}$$

$$\mathrm{Log}\{p(\vec{X} \mid w_k)\} = -\frac{D}{2}\mathrm{Log}\{2\pi\} - \frac{1}{2}\mathrm{Log}\{\det(C_k)\} - \frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1}(\vec{X}-\vec{\mu}_k)$$

One note que $-\frac{D}{2}\mathrm{Log}\{2\pi\}$ peut être eliminé parce qu'il est constant pour tout k.

La fonction de discrimination devient :

$$g_k(\vec{x}) = -\frac{1}{2}\,\mathrm{Log}\{\det(C_k)\} - \frac{1}{2}(\vec{X} - \vec{\mu}_k)^T C_k^{-1}(\vec{X} - \vec{\mu}_k) + \mathrm{Log}\{p(w_k)\}$$

Les classifiers Bayesiennes sont définies par les variations de ce formula.

Forme Cannonique de la fonction de descrimination

La classe w_k est celle qui donne un maximum pour

$$g_k(\vec{X}) = -\frac{1}{2}(\vec{X} - \vec{\mu}_k)^T C_k^{-1}(\vec{X} - \vec{\mu}_k) + \frac{1}{2}\,\mathrm{Log}\{\det(C_k)\} + \mathrm{Log}\{p(w_k)\}$$

On peut réécrire $(\vec{X} - \vec{\mu}_k)^T C_k^{-1}(\vec{X} - \vec{\mu}_k)$ comme

$$\vec{X}\,C_k^{-1}\vec{X} - \vec{X}\,C_k^{-1}\vec{\mu}_k - \vec{\mu}_k^T C_k^{-1}\vec{X} + \vec{\mu}_k^T C_k^{-1}\,\vec{\mu}_k$$

On note que C_k^{-1} est symmetrique, et donc $\vec{X}\,C_k^{-1}\vec{\mu}_k = \vec{\mu}_k^T C_k^{-1} X$

Donc $-\vec{X}\,C_k^{-1}\vec{\mu}_k - \vec{\mu}_k^T C_k^{-1}\vec{X} = 2\,(\vec{\mu}_k^T C_k^{-1})^T\,\vec{X} = 2\,(C_k^{-1}\vec{\mu}_k)^T\,\vec{X}$

On peut réécrire $g_k(\vec{X})$ comme

$$g_k(\vec{X}) = -\vec{X}^T(\tfrac{1}{2}C_k^{-1})\vec{X} + (C_k^{-1}\vec{\mu}_k)^T\vec{X} - \frac{1}{2}(\vec{\mu}_k^T C_k^{-1}\vec{\mu}_k)$$
$$-\frac{1}{2}\mathrm{Log}\{\det(C_k)\} + \mathrm{Log}\{p(w_k)\})$$

ou bien

$$g_k(\vec{X}) = \vec{X}^T(D_k)\,\vec{X} + \vec{d}_k^{\,T}\,\vec{X} + d_{ko}.$$

avec $D_k = \frac{1}{2}C_k^{-1}$ $\qquad \vec{d}_k = C_k^{-1}\vec{\mu}_k$

$$d_{ko} = -\frac{1}{2}(\vec{\mu}_k^T C_k^{-1}\vec{\mu}_k) - \frac{1}{2}\mathrm{Log}\{\det(C_k)\} + \mathrm{Log}\{p(w_k)\}$$

Cette fonction est composé de trois termes :

une terme quadrique $\vec{X}^T(D_k)\,\vec{X}$m,

une terme linéaire : $\vec{d}_k^{\,T}\,\vec{X}$

et une terme constant : d_{ko}