Systèmes Intelligents : Raisonnement et Reconnaissance

James L. Crowley

Deuxième Année ENSIMAG Deuxième Semestre 2005/2006

Séance 9 5 avril 2006

Reconnaissance Bayesienne

Notations 2

La Loi Normale.....3

Estimations des moments d'une densité.....4

Le premier moment : La Moyenne.....4

Le deuxième moment (La variance).....5

La Loi Normale pour D = 1.....6

La Loi Normale pour D > 1.....7

Forme en Algebre Linéaire.....12

Transformations Linéaire.....13

Fonctions de Discrimination 14

Discrimination.....14

Bruit d'observation.....17

Classification pour K > 2 et D > 1.....18

Forme Canonique de la fonction de discrimination.....20

Bruit et Choix de la Fonction de Discrimination.....21

Formes Quadratique.....22

Cas des classes avec moyennes égales.....24

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

M_k	Nombre d'exemples de la classe k.
M	Nombre totale d'exemples de toutes les classes

$$M = \sum_{k=1}^{K} M_k$$

ω_k	L'affirmation que Evennement E T_k

$h(x)$	Histogrammes des valeurs (x est entieres avec range limité)
$h_k(x)$	Histogramme des valeurs pour la class k.

$$h(x) = \sum_{k=1}^{K} h_k(x)$$

ω_k	Proposition que l'evennement E la classe k
$p(\omega_k) = p(E \quad T_k)$	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

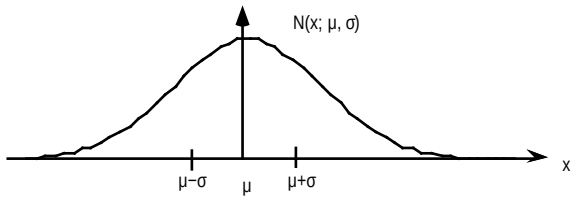
$$P(\vec{X}) = \sum_{k=1}^{K} p(\vec{X} \mid \omega_k)\, p(\omega_k)$$

La 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.

Soit M exemple d'observation d'une Evennement E_m : X_m
La loi Normale décrit une population d'exemples $\{X_m\}$.
Les paramètres de $\mathcal{N}(\mu, \sigma)$ sont les premiers et deuxième moments de la population.
On peut estimer les moments pour n'importe quel nombre d'exemples (M > 0)

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



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 difference entre x et μ normalisée par la variance.
La différence $(x - \mu)^2$ est la "distance" entre une caractéristique et la moyenne de l'ensemble $\{X_m\}$. 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, $\{X_m\} : \{ X_1, X_2, ..., X_M\}$
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).
On note que dans le cas ou il existe une histogramme pour X, 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 $[x_{min}, x_{max}]$ on peut démontrer que

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

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 la moyenne peut être calculé par le 1e moment du pdf.

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

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 d'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 !) Ils ont les usages différents.

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, la variance est la deuxième moment de la pdf.

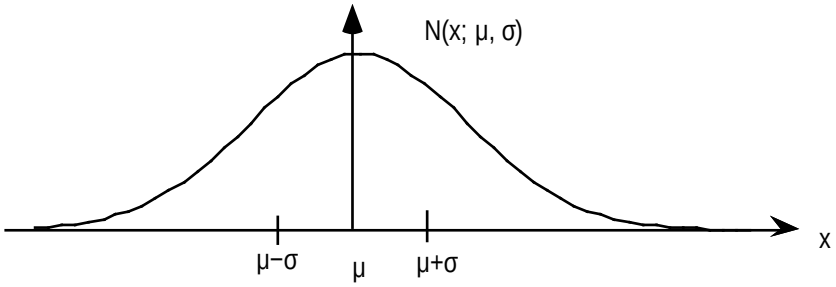
$$\sigma^2 \equiv E\{(X_m-\mu)^2\} = \int p(x)\cdot(x-\mu)^2\,dx$$

La Loi Normale pour D = 1

La cas le plus simple concerne une seule caractéristique.
Avec μ et σ^2, on peut estimer la densité p(x) par $\mathcal{N}(x; \mu, \sigma)$

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

$\mathcal{N}(x; \mu, \sigma)$ 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

Soit les événements E décrit par une vecteur de D caractéristiques $\vec{X}$

Soit une ensemble de M événements, $\{E_m\}$ avec leurs caractéristiques. $\{\vec{X}_m\}$
Cet ensemble est dit l'ensemble d'entrainement (training set)

$$\mu_d \equiv E\{x_d\} = \frac{1}{M}\sum_{m=1}^{M} X_{dm}$$

Pour le vecteur de D caractéristiques :

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

Pour M observations $\{\vec{\mathbf{X}}_m\}$, la covariance entre les variables x_i et x_j est

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

Ces coefficients composent une matrice de covariance. $\mathbf{C}_x$

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

En matrice en écrit :

Soit $\vec{V} = \vec{X} - E\{\vec{X}\} = \vec{X} - \vec{\mu}$

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

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

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

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

Dans tous les cas :

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

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

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

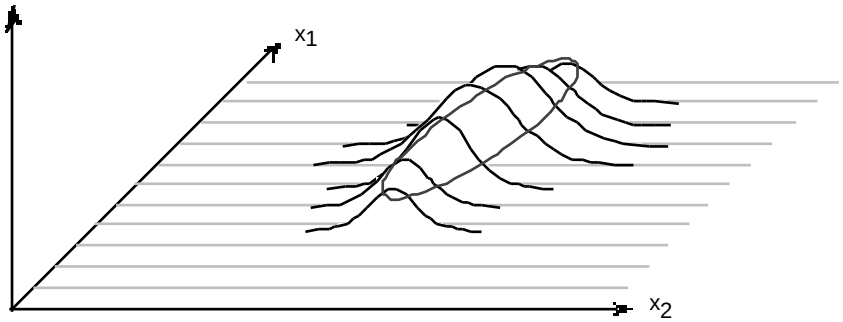
Ces coefficients composent une matrice de covariance. C

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

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

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

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



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

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

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$^{2\cdot}$)

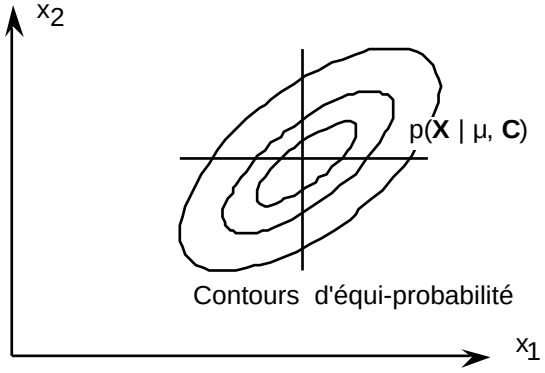
Cette valeur est connue comme la "distance de Mahalanobis".

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

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

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

Soit les événements E décrit par une vecteur de caractéristiques $\vec{X}$: $(E,\vec{X})$.
Soit une ensemble aléatoire de M événements avec leurs caractéristiques.
Cet ensemble est dit l'ensemble d'entrainement (training set) $\{\vec{X}_m\}$

Pour un vecteur de D caractéristiques :

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

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 $\{\vec{\mathbf{X}}_m\}$

Soit $\vec{V} = \vec{X} - E\{\vec{X}\} = \vec{X} - \vec{\mu}$

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

Ces coefficients composent une matrice de covariance. $\mathbf{C}_x$

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

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

Forme en Algebre Linéaire

Une expression en algebre de Matrice est souvent utile.

Soit une ensemble aléatoire de M événements avec leurs caractéristiques. $\{\vec{X}_m\}$

$$\vec{\mu} \equiv E\{\vec{X}\} = \frac{1}{M} \sum_{m=1}^{M} \vec{X}_m$$

Soit $\vec{V}_m = \vec{X}_m - \vec{\mu}$

On peut faire une matrix V composé de M colones $\{\vec{V}_m\}$

$$V = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1M} \\ v_{21} & v_{22} & \cdots & v_{2M} \\ \cdots & \cdots & & \cdots \\ v_{D1} & v_{D2} & \cdots & v_{DM} \end{pmatrix}$$

$$\mathbf{C}_X \equiv V\ V^T = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

$$= \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

Transformations Linéaire

La transformation linéaire d'une loi normale et une loi normale. Les moments d'une transformation linéaire sont les transformations linéaires des moments.

Soit un vecteur unitaire $\vec{R} = \begin{pmatrix} \Delta x_1 \\ \Delta x_2 \\ \dots \\ \Delta x_n \end{pmatrix} = \begin{pmatrix} \cos(\alpha_1) \\ \cos(\alpha_2) \\ \dots \\ \cos(\alpha_D) \end{pmatrix}$ tel que $\| \vec{R} \| = 1$.

La projection (transformation linéaire) de $\vec{X}$ sur y est

$$y = \vec{R}^T \vec{X}.$$

Pour la covariance :

$$\begin{aligned} \sigma_y^2 &\equiv E\{(\vec{R}^T \vec{V})(\vec{R}^T \vec{V})^T\} \\ &= E\{(\vec{R}^T \vec{V})(\vec{V}^T \vec{R})\} \qquad \text{car } (\vec{R}^T \vec{V})^T = (\vec{V}^T \vec{R}) \\ &= E\{\vec{R}^T (\vec{V} \vec{V}^T) \vec{R}\} \\ &= \vec{R}^T E\{\vec{V} \vec{V}^T\} \vec{R} = \vec{R}^T \mathbf{C}_x \vec{R} \end{aligned}$$

La projection de la covariance est la covariance de la projection.
La projection de la moyenne et la covariance sur un axe, $\vec{R}^T$ donne une moyenne μ_y et variance, σ_y^2 dans la direction $\vec{R}$.

$$\mu_y = \vec{R}^T \vec{\mu}_x, \qquad \sigma_y^2 = \vec{R}^T \mathbf{C}_x \vec{R}$$

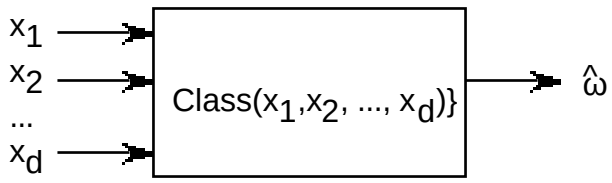
$$p(y) = \mathcal{N}(y; \vec{R}^T \vec{\mu}_x, \vec{R}^T \mathbf{C}_x \vec{R}^T) = \mathcal{N}(y; \mu_y, \sigma_y^2)$$

Les moments d'une projection sont les projections des moments.

$$\mu_y = E\{p(y)\} = \vec{R}^T \vec{\mu}_x \qquad \sigma_y^2 = E\{(p(y)-\mu_y)(p(y)-\mu_y)\} = \vec{R}^T \mathbf{C}_x \vec{R}$$

Fonctions de Discrimination

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



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

$\hat{\omega}_k$ est la proposition que (E ω_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 : $R^K \rightarrow \{\omega_K\}$

Discrimination

$g(\vec{X})$: Une fonction de discrimination est une fonction $R^D \rightarrow R^K$

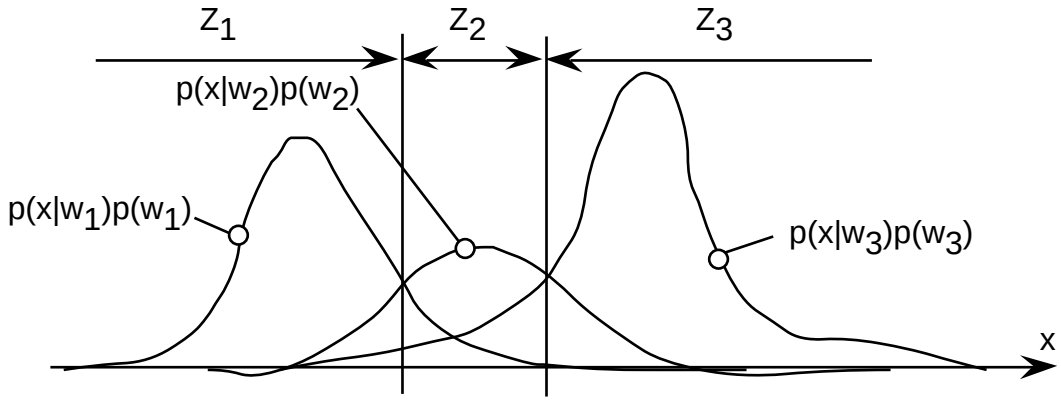
$$g(\vec{X}) = \begin{pmatrix} g_1(\vec{X}) \\ g_2(\vec{X}) \\ \dots \\ g_K(\vec{X}) \end{pmatrix}$$

Dans le cas général, K > 2, la nombre minimum d'erreur sont fait si ω_k est choisi tel que :

$k = \arg\max_k \{g_k(X)\}$ avec $g_k(X) = p(x \mid \omega_k)\, p(\omega_k)$

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, ..., Z_k pour chaque classe.

$$k = \underset{k}{\text{arg-max}}\{g_k(X)\}$$

Mais comment calculer $g_k(X)$?

Les caractéristiques X de l'événement E sont aléatoires avec une dispersion due aux variations naturelles de sa classe.

Ceci est modélisé par une variable aléatoire B_k autours d'une valeur "type" x_k. La valeur type est spécifique à la classe.

$$X = x_k + B_k$$

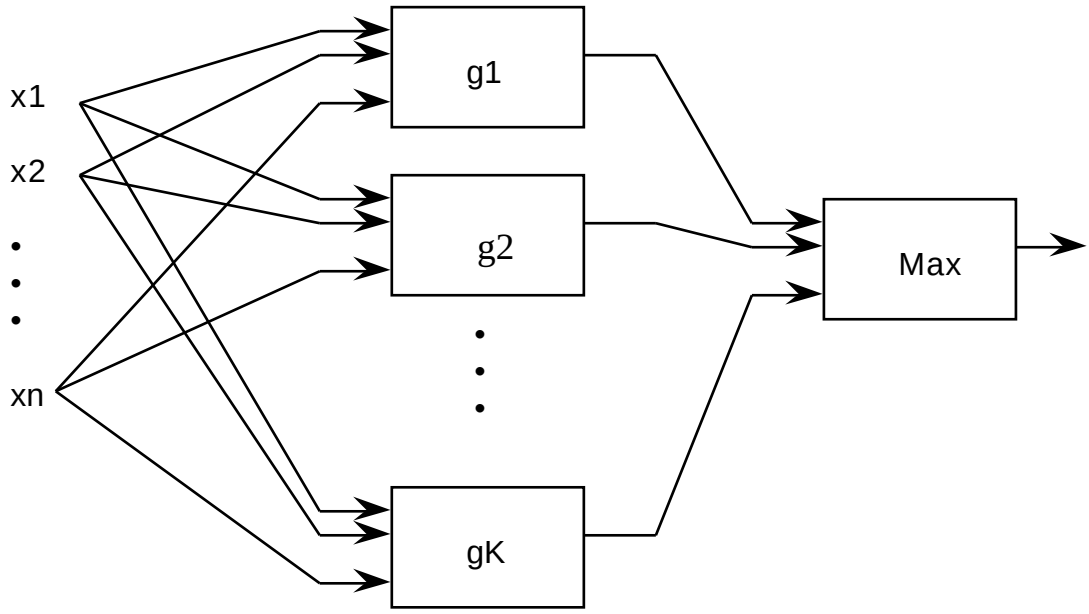
Si D=1, les membres de la classe ω_k auront les caractéristiques X tel que :

$$p(X=x \mid \omega_k) = \mathcal{N}(x; \mu_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

Donc notre fonction de discrimination devient :

$$g_k(X) = p(\omega_k) \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

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



On peut noter que $k = \arg\max_k\{g_k(X)\} = \arg\max_{\omega_k}\{\mathrm{Log}\{g_k(X)\}\}$

parce que Log{} est une fonction monotone.

$$k = \arg\max_k\{\mathrm{Log}\{\frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}\} + \mathrm{Log}\{p(\omega_k)\}\}$$

$$k = \arg\max_k\{\mathrm{Log}\{\frac{1}{\sqrt{2\pi}\sigma_k}\} + \mathrm{Log}\{e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}\} + \mathrm{Log}\{p(\omega_k)\}\}$$

$$k = \arg\max_k\{-\mathrm{Log}\{\sqrt{2\pi}\ \sigma_k\} - \frac{(x-\mu_k)^2}{2\sigma_k^2} + \mathrm{Log}\{p(\omega_k)\}\}$$

$$k = \arg\max_k\{-\mathrm{Log}\{\sigma_k\} - \frac{(x-\mu_k)^2}{2\sigma_k^2} + \mathrm{Log}\{p(\omega_k)\}\}$$

Bruit d'observation

Chaque observation d'un événement corrompu par une bruit d'observation.

$$Y = x_k + B_k + B_o$$

B_o est souvent Normale, avec moyenne 0 et variance σ_o^2.
On dit que la variance σ_o^2 est la précision de la capteur.
Cette bruit ne depend pas de la classe.

Dans ce cas, on observe Y avec

$$p(Y=y \mid \omega_k) = \mathcal{N}(y; \mu_k, \sigma_k^2 + \sigma_o^2) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\frac{(y-\mu_k)^2}{(\sigma_k^2+\sigma_o^2)}}$$

et

$$k = \arg\max_k \{-\mathrm{Log}\{\sqrt{\sigma_k^2+\sigma_o^2}\} - \frac{(y-\mu_k)^2}{2(\sigma_k^2+\sigma_o^2)} + \mathrm{Log}\{p(\omega_k)\}\}$$

$$= \arg\max_k \{-\mathrm{Log}\{\sigma_k\} - \frac{(y-\mu_k)^2}{2(\sigma_k^2+\sigma_o^2)} + \mathrm{Log}\{p(\omega_k)\}\}$$

Classification pour K > 2 et D > 1.

Dans le cas général, il y a D caractéristique.

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

Et le règle de décision est :

$$\hat{\omega}_i \; : \text{si} \quad j \neq i \quad g_i(\vec{X}) > g_j(\vec{X})$$

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

La fonction de discrimination est : $g_k(\vec{X}) = p(\vec{X} \mid \omega_k) \, p(\omega_k)$

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

par règle de Bayes :

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

$$= \arg\max_k \{\text{Log}\{p(\vec{X} \mid \omega_k)\} + \text{Log}\{p(\omega_k)\}$$

Si les caractéristiques suivent une densité Normale :

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

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

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

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

La fonction de discrimination devient :

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

Les classifiers Bayesiennes sont définies par les variations de cette formule.

Forme Canonique de la fonction de discrimination

La décision ω_k est celle qui donne un maximum pour

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

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 symétrique, 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

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

ou bien

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

avec
$$D_k = \frac{1}{2}\, C_k^{-1}$$
$$\vec{d}_k = C_k^{-1}\vec{\mu}_k$$
$$d_{ko} = -\frac{1}{2}(\vec{\mu}_k^T C_k^{-1}\vec{\mu}_k) - \frac{1}{2}\, \text{Log}\{\det(C_k)\} + \text{Log}\{p(\omega_k)\}$$

Cette fonction est composée de trois termes :

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

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

et une terme constant : d_{ko}

Bruit et Choix de la Fonction de Discrimination

La conception d'un classifier dépend de la choix de caractéristiques, $\vec{x}$ et du bruit observé sur ses caractéristiques :

Rappel qu'une observation $\vec{Y} = \vec{x}_k + \vec{B}_k + \vec{B}_o$

où

$\vec{x}_k$ Est la forme type (moyenne) de la classe w_k

$\vec{B}_k$: Les variations aléatoires intra-classe.
Elle est spécifiques à chaque classe et chaque individus.
Elles n'est change pas entre observations.

$\vec{B}_o$: Les variations aléatoires des observations.
Elles est indépendantes de la classe et de l'individu.
Elles changent avec les observations.

Selon la nature de $\vec{B}_k$, $\vec{B}_o$ et de $p(\omega_k)$ on peut faire certaines simplifications.

par exemple :

Si $\vec{B}_o >> \vec{B}_k$ on peut trouver que w_k : $C_k \approx C$: On a une classifier linéaire.

Si i,j $\sigma_{ij}^2 = \sigma^2$ C : On a la détecteur optimale utilisé en communication hz.

Exemples de caractéristiques : $\vec{x}$

1) Les échantillons d'un signal : x(n) pour n [1, N]
2) Les caractéristiques d'un individu : [hauteur, poids, yeux, cheveux etc.]
3) Les caractéristiques géométriques d'un objet : Hauteur, largeur, nombre de faces, etc.

Formes Quadratique

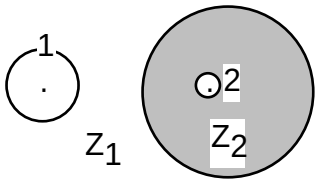
Dans le cas le plus général, on ne fait aucune hypothèse sur B_k et $\vec{B}_o$
Dans ce cas, ω_k, C_k est arbitraire.

La surface de décision est une fonction quadrique en D dimensions.
(une hyper quadriques)

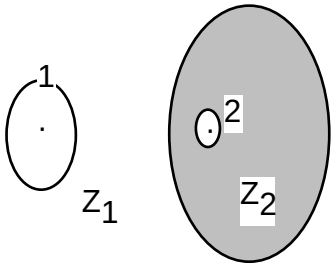
Elle peut être les hyperplans, hyper-sphéres, hyper-ellipsoides, hyper-parabaloides, ou les hyperhyperboloides.

Par exemples, en 2D (D=2) quand K = 2.

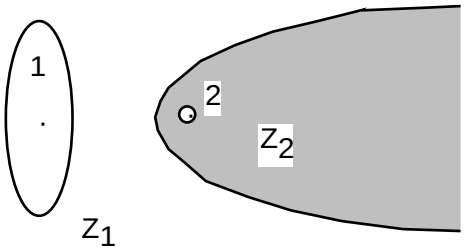
Hyper-sphère :
Pour k = 1, 2 $C_k = \sigma_k^2 I$
et $\det\{C_1\} > \det\{C_2\}$



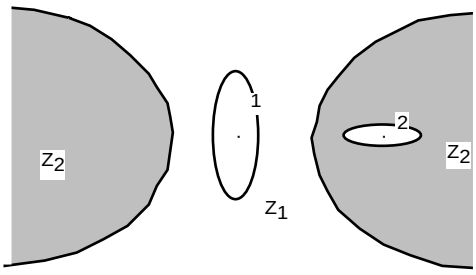
Hyper-éllipsoide :
Pour k = 1, 2 $\sigma_{x1}^2 > \sigma_{x2}^2$
et $\det\{C_1\} > \det\{C_2\}$



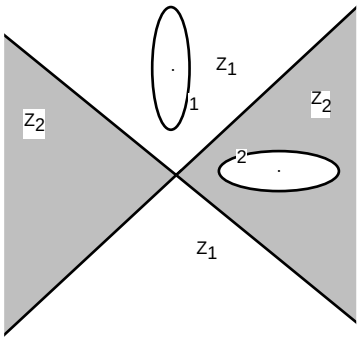
Hyper-paraboloide :
Pour k = 1, 2 $\sigma_{x1k=1}^2 >> \sigma_{x1k=2}^2$
et $\sigma_{x2k=1}^2 > \sigma_{x2k=2}^2$



Hyper-hyperboloides :

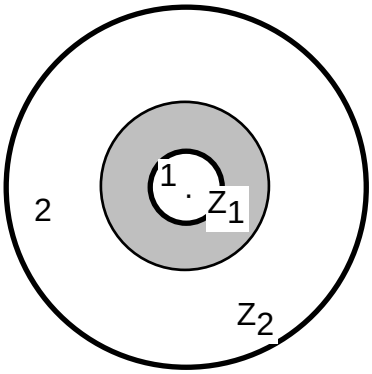


Hyperplanes.

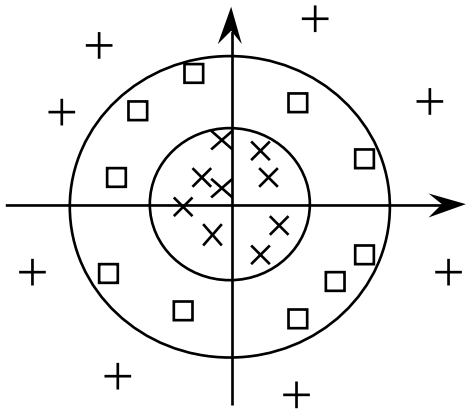


$\vec{\mu}_1 = \vec{\mu}_2$ et $C_1 << C_2$
avec $\sigma_{11} = \sigma_{22}$ et $\sigma_{12} = \sigma_{21} = 0$.

Une hypershere.



Cas des classes avec moyennes égales.



Supposons que nous avons K classes tel que

$$i, j : \quad \vec{\mu}_i = \vec{\mu}_j \text{ et } \det(C_i) \neq \det(C_j).$$

Comment peut on decider la classe d'un evennement $(E, \vec{X})$?

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

1) $D_k = \frac{1}{2} C_k^{-1}$ est discriminant.

2) $\vec{d}_k = C_k^{-1} \vec{\mu}_k = C_k^{-1} \vec{\mu}$ peut être éliminé.

3) $d_{ko} = -\frac{1}{2} \vec{\mu}_k{}^T C_k^{-1} \vec{\mu}_k + \text{Log}\{p(w_k)\}$ est réduit à

$$d_{ko} = -\frac{1}{2} \vec{\mu}^T C_k^{-1} \vec{\mu} + \text{Log}\{p(\omega_k)\}$$

Il s'agit d'un biais pour chaque classe

donc : $g_k(\vec{X}) = \vec{X}^T (\frac{1}{2} C_k^{-1}) \vec{X} + \text{Log}\{p(\omega_k)\}$

.

Les surfaces de décisions entre classes i et j sont les hyper-surfaces

telles que $g_i(\vec{X}) - g_j(\vec{X}) = 0$ sont les hyper surfaces.