Systèmes Intelligents : Raisonnement et Reconnaissance

James L. Crowley

Deuxième Année ENSIMAG Deuxième Semestre 2008/2009

Séance 9 29 avril 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

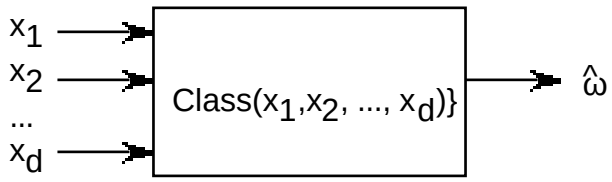
x	Une variable
X	Une valeur aléatoire (non-prévisible).
N	Le nombre de valeurs possible pour x ou X
$\vec{x}$	Un vecteur de D variables
$\vec{X}$	Un vecteur aléatoire (non-prévisible).
D	Nombre de dimensions de $\vec{x}$ ou $\vec{X}$
E	Une événement, un entité
A,B	des classes d'événements ou d'entités
T_k	La classe k d'événements ou d'entités
k	Indice d'une classe
K	Nombre de classes
ω_k	La proposition que l'entité 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

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

La classification est un processus d'estimation de l'appartenance de l'entité E 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 :

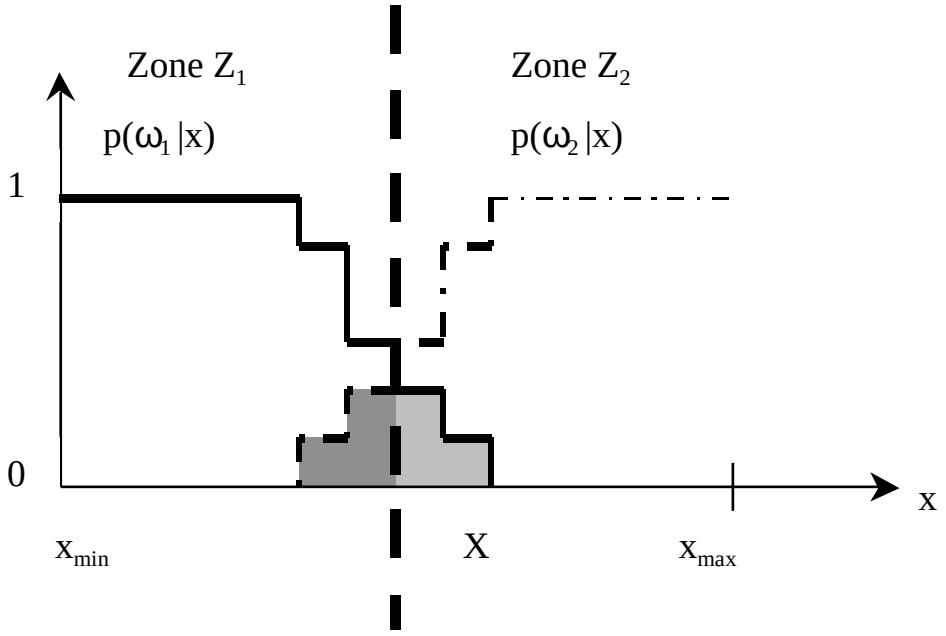
$$p(\omega_k) = p(E \quad T_k)$$

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

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

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

ω_1 si X Z_1 et ω_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.

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

La minimum est atteint quand :

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

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

Les Variations Aléatoires des Caractéristiques

Les vecteurs de caractéristique porte une composante "aléatoire" avec (au moins) deux origines :

1) Les variations des individus.
2) Le bruit des capteurs

1) Les variations entre individus d'une classe

La formation des vrais objets physiques est asujetté aux influences aléatoires.
Pour les objets d'une classe, T_k, les propriétés des objets individuels sont,
les valeurs aléatoires. On peut résume 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 plupart de techniques probabilistes de reconnaissance supposent un bruit additif.
En notation vectorielle :

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

Attention. Le bruit peut etre non-additif, et ceci devient une approximation.

2) Le bruit des capteurs

(def.) 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 : { Z_1, Z_2 ... Z_D}.

$$\vec{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \dots \\ Z_n \end{pmatrix}$$

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

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

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 ω_k, la probabilité d'observé $\vec{Z}$ est fournie par la règle de Bayes.

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

Si $\vec{Z} = F(\vec{X} + \vec{B}_o)$ est issu de la classe ω_k ayant caractéristique $\vec{X} = \vec{x} + \vec{B}_i$

Une Exemple - Le spectre observées par un satellite.

Une image satellite est composée de pixels s(i,j). Chaque pixel compte le nombre de photons issus d'une surface carré de la terre (ex. 10 m^2·).

Les photons sont captés au travers des filtres spectraux. Ceci donne un vecteur de caractéristiques pour chaque pixel. Soit une région de végétation {blé, maize, etc}

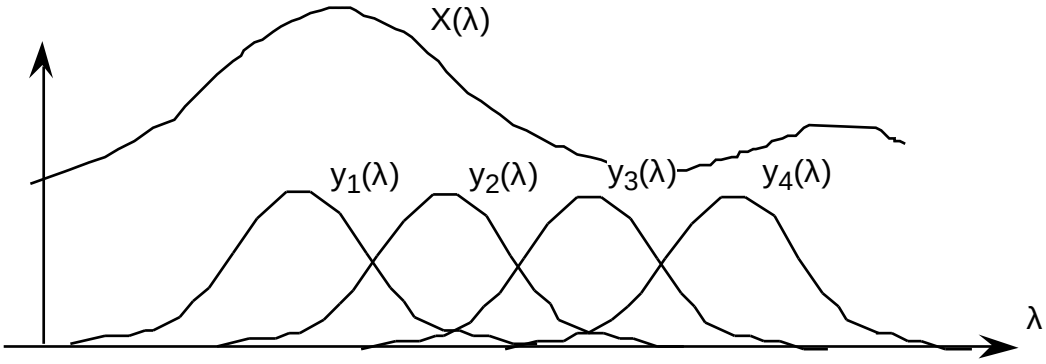
$\vec{x}$: Le spectre des pigments des feuilles pour une espèce.

$\vec{B}_i$: Les variations du spectre intrinsèque dû aux variations d'age ou d'humidité.

$\vec{B}_i$ est spécifique à un individu. Il ne change pas entre les observations.

$\vec{X} = \vec{x} + \vec{B}_i$: Le spectre des pigments des feuilles pour un individu

$\vec{B}_o$: Les variations d'observations dues à l'angle du soleil et les effets de filtrage de la lumière par l'atmosphère (humidité, pollution etc). Ils sont présents dans tous les pixels.



Une image est une table de pixels, avec les positions sur les lignes (j) et colonne (i). Un pixel, $\vec{Z}(i,j)$, est un vecteur d'entiers, Z_1, ..., Z_D. Chaque composant est l'intégral sur i,j et λ (longueurs d'onde) d'une produit d'un filtre spectral avec le spectre reçu sur la région c,r.

$$Z_d(i,j) = \text{Quant}\{ \int_i^{i+di} \int_j^{j+dj} \int_\lambda^{\lambda+\delta\lambda} X(i,j,\lambda) \cdot f_d(\lambda),\ di\ dj\ d\lambda. \}$$

C'est opération est réalisé par l'optique de la caméra. L'opération "Quant{} numérise les valeurs y_d avec un pas "q" sur une plage entre 0 et v_{max}.

Une classification est une estimation de la classe de végétations dominant, ω_k, pour la région observée par le pixel à partir de le vecteur d'observation $\vec{Z}$

$$\hat{\omega}_k \ ?= \text{Class}\{ \vec{Z} \}$$

La Loi Normale

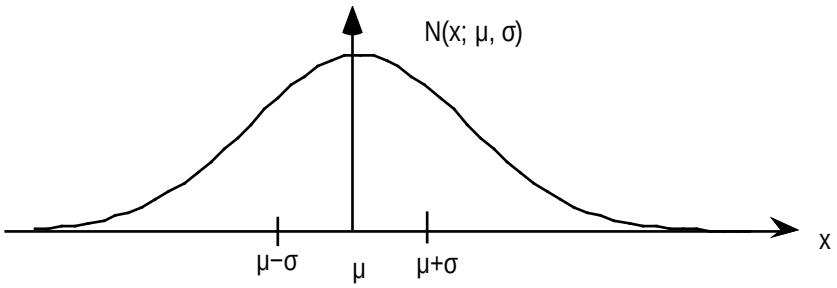
La fonction paramétrique la plus utilisée est la loi Normale.

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.

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)
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 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.

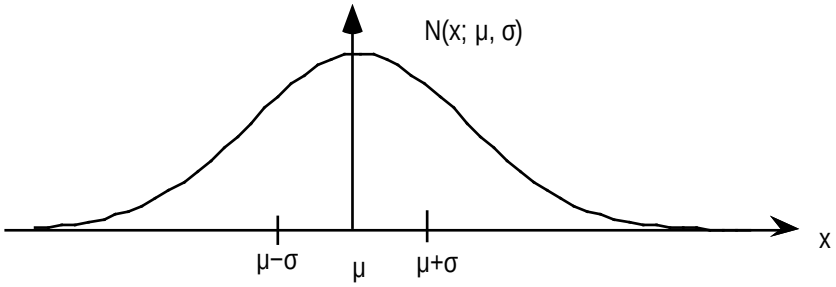
La différence normalisée par la variance est connue sous le nom de "Distance de Mahalanobis". La Distance de Mahalanobis est un test naturel de similarité

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

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

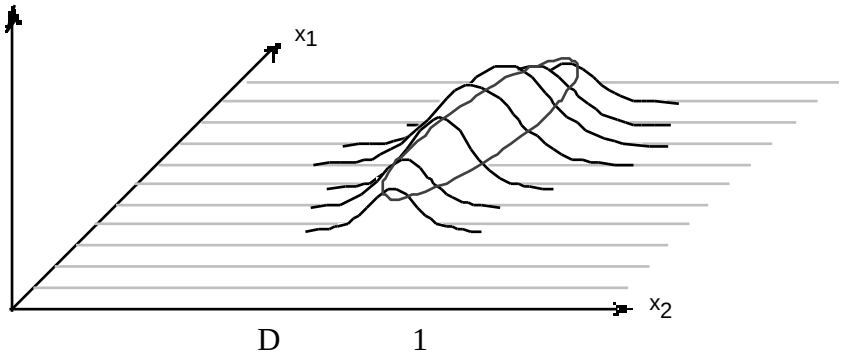
Ces coefficients composent une matrice de covariance. C

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

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

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$^{2\cdot}$)

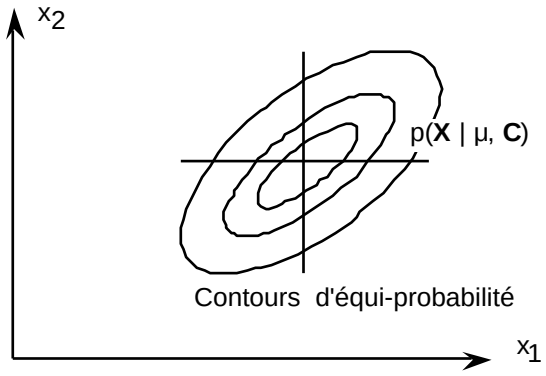
Cette valeur est connue comme la "distance de Mahalanobis".

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

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

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

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

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

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

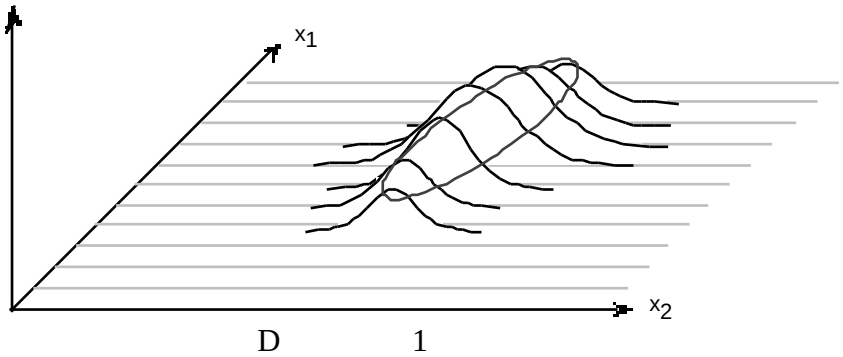
Ces coefficients composent une matrice de covariance. C

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

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

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

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



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

Forme en Algebre Linéaire

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 :

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

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

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

Ceci peut etre exprimé en forme de matrice.

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

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

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

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

$\mathbf{C}_X \equiv V\,V^T$ est D x D. Note que $\mathbf{C}_m = V^T V$ est de taille M x M.

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 $\vec{R}$.

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

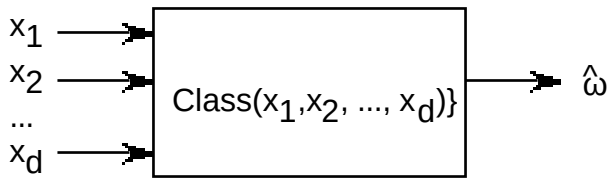
$$p(y) = \mathcal{N}(y; \vec{R}^T \vec{\mu}_x, \vec{R}^T \mathbf{C}_x \vec{R}^T) = \mathcal{N}(y; \mu_y, \sigma_y^2)$$

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 $R^D \rightarrow R^K$

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

Etant donnée $\vec{X}$, pour chaque k il existe une valeur de probabilité $p(\omega_k \mid X)$

$$p(\omega_k \mid \vec{X}) = \frac{P(\vec{X}|\omega_k)}{P(\vec{X})} \; p(\omega_k)$$

Dans le cas général la nombre minimum d'erreur est fait si ω_k est choisi tel que

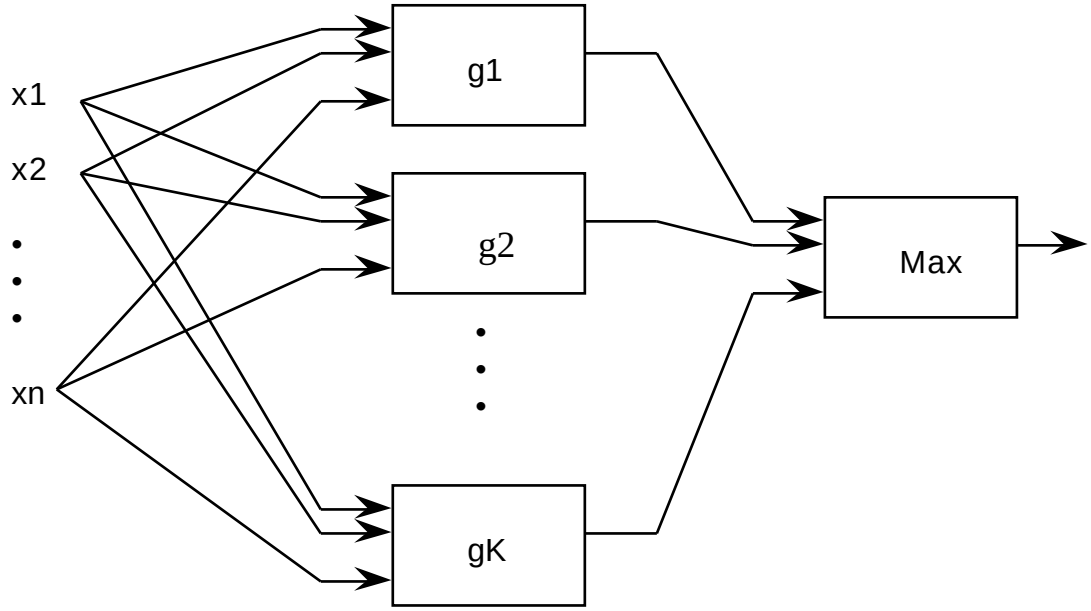
$$k = \underset{k}{\text{arg-max}}\{g_k(X)\} = \underset{k}{\text{arg-max}}\{p(\omega_k \mid \vec{X})\} = \underset{k}{\text{arg-max}}\left\{\frac{P(\vec{X}|\omega_k)}{P(\vec{X})}\ p(\omega_k)\right\}$$

mais, comme $P(\vec{X})$ est constant pour tous k,

$$k = \underset{k}{\text{arg-max}}\{\ (P(\vec{X}|\omega_k)\ p(\omega_k)\}$$

Il suffit de l'evaluer $P(\vec{X}|\omega_k)$, pour $\vec{X}=\vec{x}$

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



Fonctions classiques :

$$P(\vec{X}|\omega_k) = \mathcal{N}(\vec{X};\ \vec{\mu}_k\ ,\mathbf{C}_k)$$

ou encore

$$P(\vec{X}|\omega_k) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(\vec{X};\ \vec{\mu}_{kn}\ ,\mathbf{C}_{kn})$$

Simplification de la fonction de Discrimination

Soit D=1, avec

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

Donc notre fonction de discrimination devient :

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

On peut noter que $k = \underset{k}{\text{arg-max}}\{g_k(X)\} = \underset{\omega_k}{\text{arg-max}}\{\text{Log}\{g_k(X)\}\}$

parce que Log{} est une fonction monotone.

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

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

$$k = \underset{k}{\text{arg-max}}\{-\text{Log}\{\sqrt{2\pi} \; \sigma_k\} \; -\frac{(x-\mu_k)^2}{2\sigma_k^2} + \text{Log}\{p(\omega_k)\} \}$$

$$k = \underset{k}{\text{arg-max}}\{-\text{Log}\{\sigma_k\} \; -\frac{(x-\mu_k)^2}{2\sigma_k^2} + \text{Log}\{p(\omega_k)\} \}$$

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

par règle de Bayes :

$$\underset{k}{\text{arg-max}}\{p(\omega_k \mid \vec{X})\} = k = \underset{k}{\text{arg-max}}\{ p(\vec{X} \mid \omega_k) \, p(\omega_k) \}$$

$$= \underset{k}{\text{arg-max}}\{\text{Log}\{p(\vec{X} \mid \omega_k)\} + \text{Log}\{p(\omega_k)\}$$

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}