Formation et Analyse d'Images

James L. Crowley

ENSIMAG 3 Premier Bimestre 2008/2009

Séance 10 9 janvier 2009

Reconnaissance Baysienne

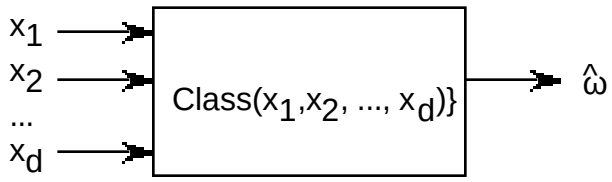
Plan de la Séance :

La Classification Bayesienne

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

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 : $R^K \rightarrow \{\omega_K\}$

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

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

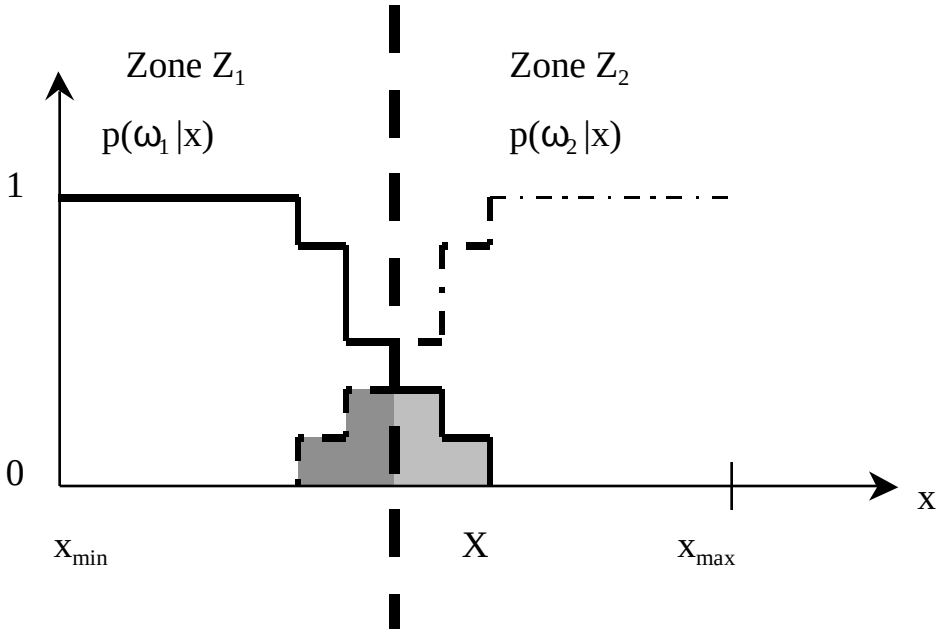
Le critère de décision est de minimiser le nombre d'erreur. Dans un système probabiliste, ca revient de minimiser la probabilité d'erreur. Ceci est équivalent à choisir la classe le plus probable.

$\hat{\omega}_k$ = Decider(E 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.

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

Pour faire $g_k(X)$ nous allons utilitser :

1) Le régle de Bayes : $$p(\omega_k \mid \vec{X}) = \frac{P(\vec{X}|\omega_k)}{P(\vec{X})}\ p(\omega_k)$$

2) La loi Normale :

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

Représentation de la Probabilité avec La Loi Normale

Une "Dénsité" de probabilité (écrit "pdf" pour "Probability Density Function")
est une fonction, g(X) represetnant la probabilité pour une variables aléateoire, X
X [–∞, ∞] tel que

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

On écrit p(X) = g(X).

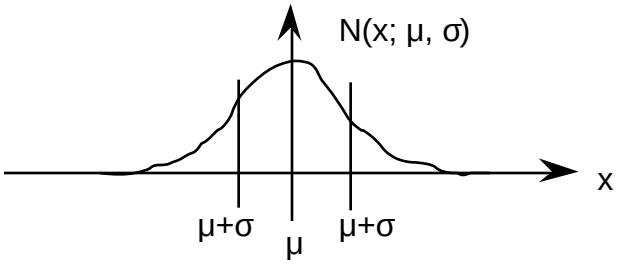
Quand les variables aléatoires sont issues d'une séquence d'événements aléatoires, leur densité de probabilité prend la forme de la loi normale, $\mathcal{N}(\mu, \sigma)$. Ceci est démontré par le théorème de la limite centrale. Il est un cas fréquent en nature.

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).

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



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 $[x_{min}, x_{max}]$ 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

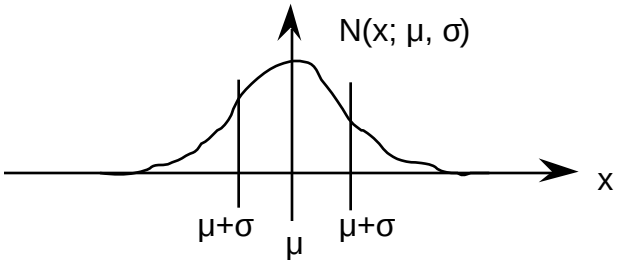
$$\sigma^2 \equiv E\{(X_m-\mu)^2\} = \int p(x)\cdot(x-\mu)^2\, dx$$

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 $\in$ [1, D], $x_d \in [x_{dmin}, x_{dmax}]$ on peut démontrer que

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

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 :

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

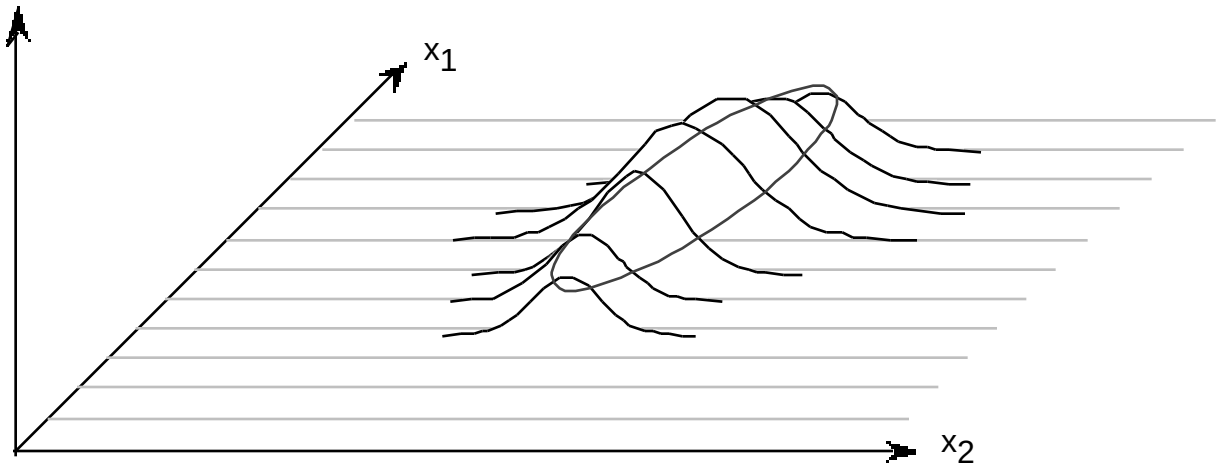
Ces coefficients composent une matrice de covariance. C

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

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

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

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

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

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$^{2\cdot}$)

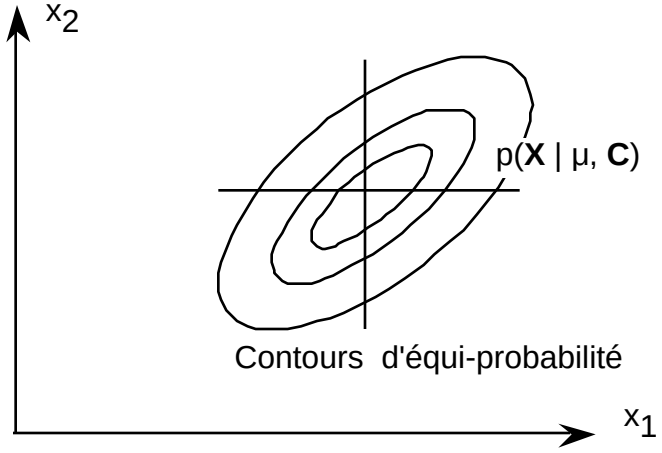
Cette valeur est connue comme la "distance de Mahalanobis".

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

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

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 $\{\vec{V}_m\}$

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

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

$\mathbf{C}_X \equiv V\ V^T$ 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 :

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

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 $\vec{R}$.

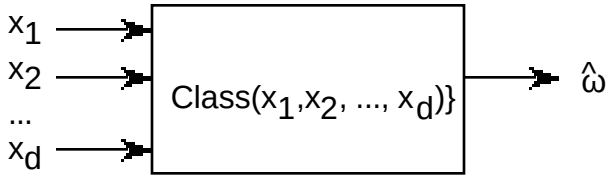
$$\mu_y = \vec{R}^T \vec{\mu}_x, \qquad \sigma_y^2 = \vec{R}^T \mathbf{C}_x \vec{R}$$

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

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 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 est une fonction $R^D \rightarrow R^K$
d() : Une fonction de décision : $R^K \rightarrow \{\omega_K\}$

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

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

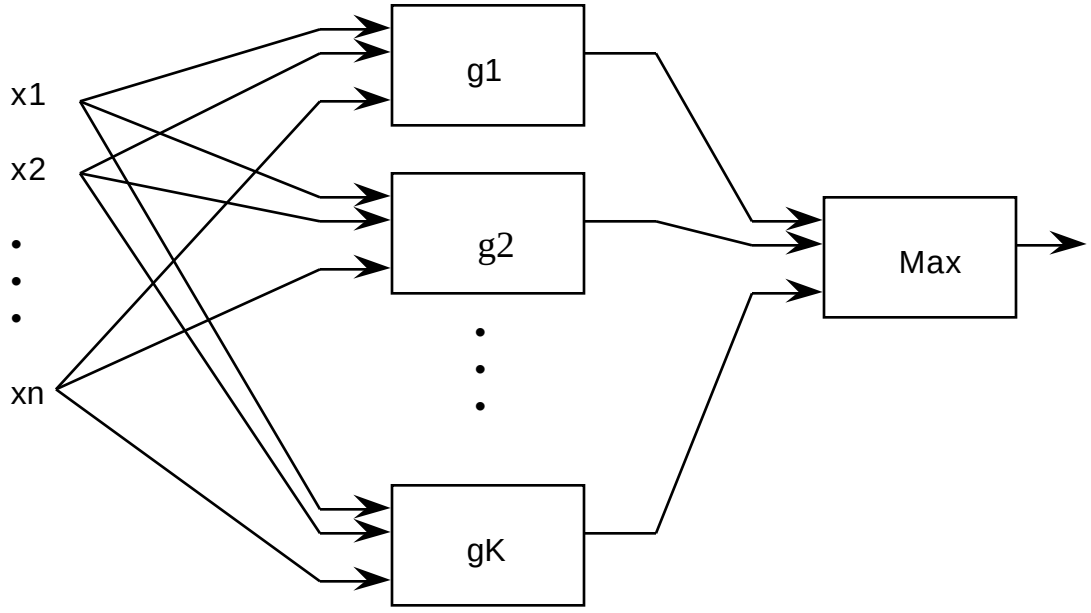
$$k = \arg\max_k\{g_k(X)\} = \arg\max_k\{p(\omega_k \mid \vec{X})\} = \arg\max_k\{\frac{P(\vec{X}|\omega_k)}{P(\vec{X})} \; p(\omega_k)\}$$

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

$$k = \arg\max_k\{ (P(\vec{X}|\omega_k) \; p(\omega_k)\}$$

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

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