Systèmes Intelligents : Raisonnement et Reconnaissance

James L. Crowley

Deuxième Année ENSIMAG Deuxième Semestre 2005/2006

Séance 8 29 mar 2006

Reconnaissance Bayesienne

Sources Bibliographiques :

"Neural Networks for Pattern Recognition", C. M. Bishop, Oxford Univ. Press, 1995.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notations

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

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 Règle de Bayes

Soit une univers d'evennement S, avec deux classes A et B. A et B ne sont pas disjoint. Soit deux propositions q et r.

$q \equiv E \quad A$ et $r \equiv E \quad B$

donc $p(q) \equiv Pr\{E \quad A\}$ et $p(r) \equiv Pr\{E \quad B\}$.

$p(q \quad r) \equiv (E \quad A) \wedge (E \quad B)$

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

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

de la même manière :

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

Par algèbre on déduire :

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

d'ou

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

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

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

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

La règle de Bayes avec une ratio d'histogrammes.

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

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

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

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

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

ce qui donne :

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

Cette technique peux également marcher pour les vecteurs de caractéristiques.

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

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

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

ce qui donne :

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

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

Exemple : Les statistiques de couleurs de la peau

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

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

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

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

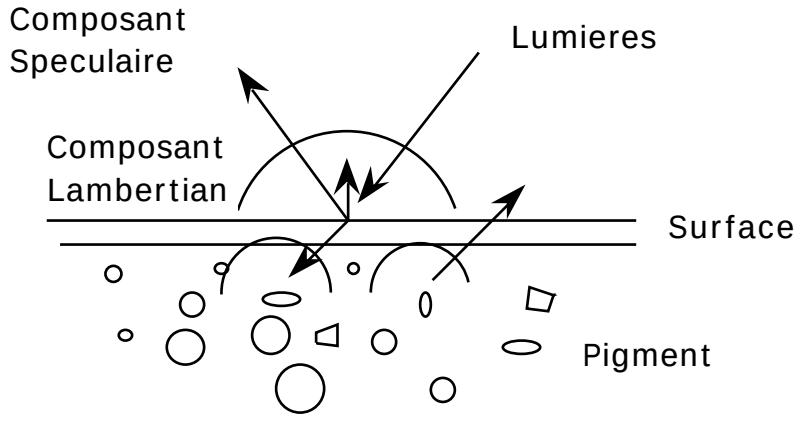
Pour un vecteur de caractéristique, on peut calculer une table à 3 dimensions.

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

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

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

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



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

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

Par exemple :

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

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

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

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

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

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

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

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

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

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

Détection par ratio d'histogramme

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

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

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

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

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

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

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

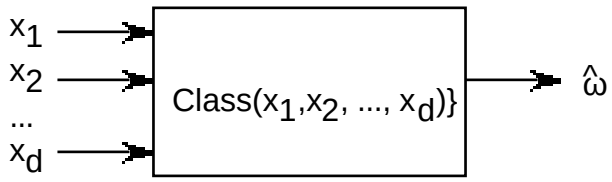
Il faut assurer que $h(\vec{x}) \neq 0$!! Pour cela il faut que $w_A(i,j)$ $w(i,j)$.

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

La Classification

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

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



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

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

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

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

La Classification Bayesienne

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

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

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

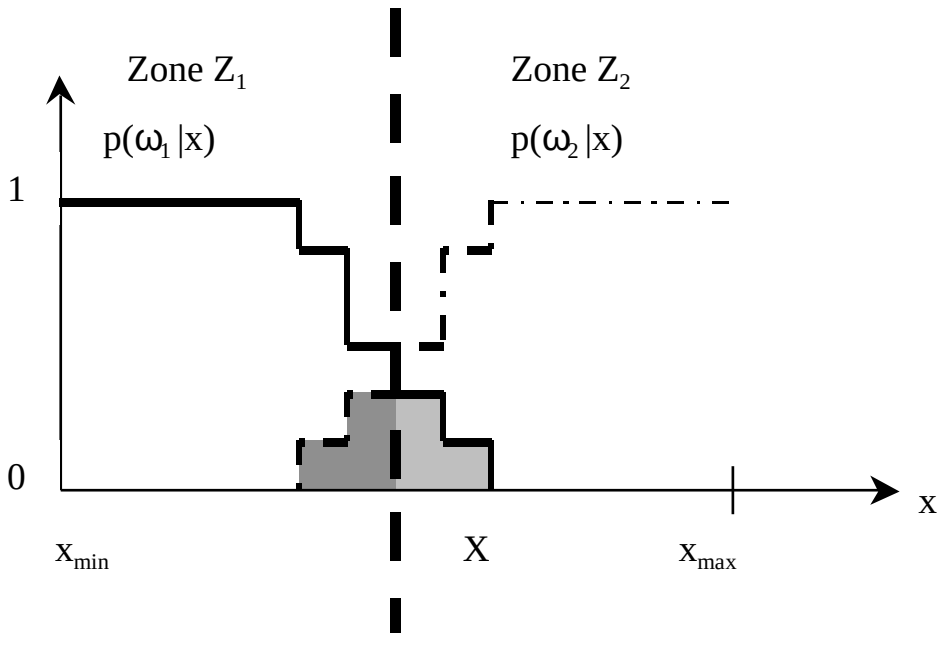
Le critère de décision est de minimiser le nombre d'erreur. Dans un système probabiliste, ca revient de minimiser la probabilité d'erreur. Ceci est équivalent à choisir la classe le plus probable.

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

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

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

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



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

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

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

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 des individus.
La formation des vrais objets physiques est sujette aux influences aléatoires.
Pour les objets d'une classe, ω_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 :

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

Ceci n'est pas strictement nécessaire. Ils existe de méthodes pour la bruit non-additif

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 : $\{ Y_1, Y_2 \ldots Y_D\}$.

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

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

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

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 $\vec{Y}$ est fournie par la règle de Bayes.

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

Si $\vec{Y} = F(\vec{X} + \vec{B}_o)$ 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(x, y). Chaque pixel compte le nombre de photons issus d'une surface carré de la terre (ex. 10 m^2.).

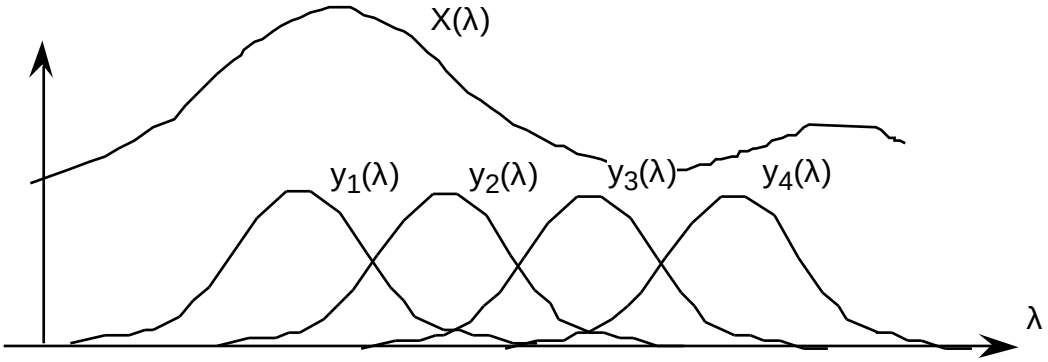
Les photons sont captés au travers des filtres spectraux. Ceci donne un vecteur de caractéristiques pour chaque pixel. Soit une région de végétation {blé, maize, etc}

$\vec{x}$: 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.



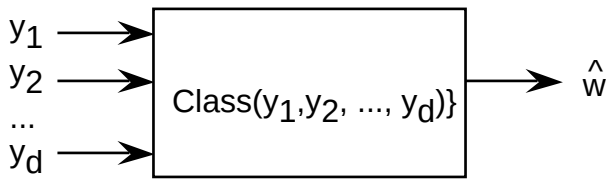
Une image est une table de pixels, avec les positions sur les lignes (r pour "row") et colonne (c). Un pixel, $\vec{Y}(c,r)$, est un vecteur d'entiers, y_1, ..., y_D. Chaque composant est l'intégral sur c, r et λ (longueurs d'onde) d'une produit d'un filtre spectral avec le spectre reçu sur la région c,r.

$$y_d(c,y) = \text{Quant}\{ \int_{c}^{c+dc} \int_{r}^{r+dr} \int_{\lambda}^{\lambda+\delta\lambda} X(c,r,\lambda) \cdot f_d(\lambda),\ dc\ dr\ d\lambda. \}$$

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 $\vec{Y}$.

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



La Loi Normale

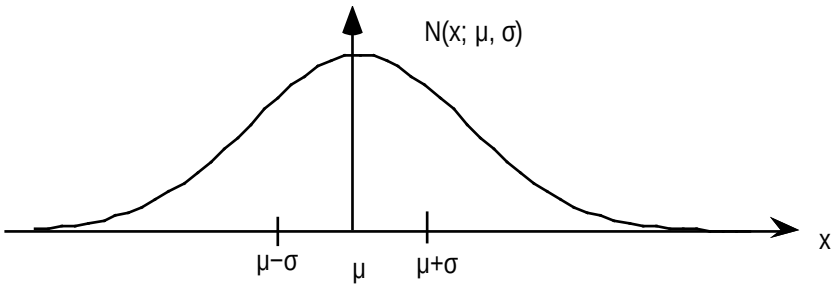
La fonction paramétrique la plus utilisée est la loi Normale.

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.

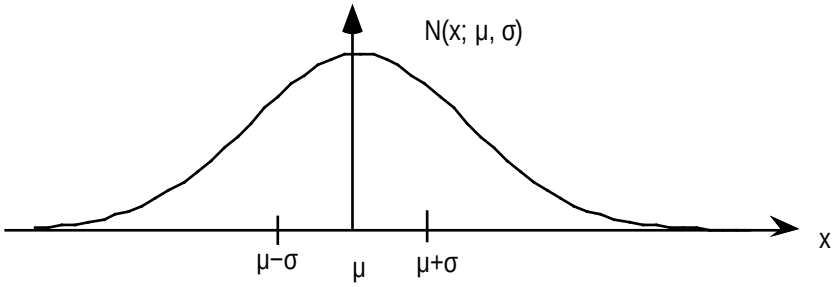
La différence normalisée par la variance est connue sous le nom de "Distance de Mahalanobis". La Distance de Mahalanobis est un test naturel de similarité

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).

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

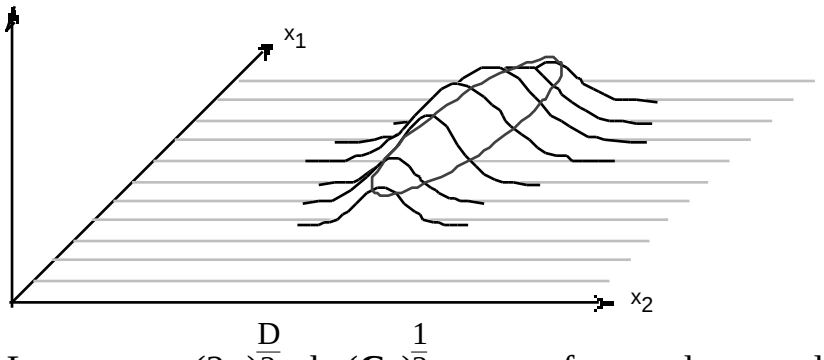
La Loi Normale pour D > 1

Pour un vecteur de D caractéristiques :

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

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.

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

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$^{2\cdot}$)

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).