Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2009/2010

Lesson 15 7 april 2010

Bayesian Recognition

Notation 2

Bayesian Classification (Reminder) 3

Normal Density Functions 4

The average value is the first moment of the samples 5

The variance is the second moment of the samples 7

Multi-Variate Normal Density Functions 8

Linear Algebraic Form for Moment Calculation 12

Linear Transforms of the Normal Multivariate Density 13

Sources Bibliographiques :

"Neural Networks for Pattern Recognition", C. M. Bishop, Oxford Univ. Press, 1995.

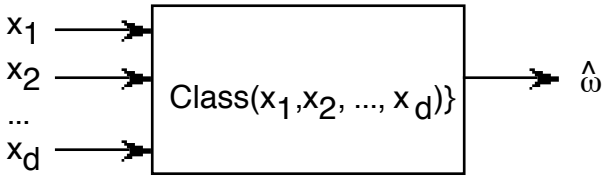
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

Notation

x	a variable
X	a random variable (unpredictable value)
N	The number of possible values for X (Can be infinite).
$\vec{x}$	A vector of D variables.
$\vec{X}$	A vector of D random variables.
D	The number of dimensions for the vector $\vec{x}$ or $\vec{X}$
E	An observation. An event.
T_k	The class (tribe) k
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that $E \in T_k$
$p(\omega_k) = p(E \in T_k)$	Probability that the observation E is a member of the class k. Note that $p(\omega_k)$ is lower case.
M_k	Number of examples for the class k. (think M = Mass)
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{X_m^k\}$	A set of M_k examples for the class k. $\{X_m\} = \bigcup_{k=1,K} \{X_m^k\}$
$P(X)$	Probability density function for X
$P(\vec{X})$	Probability density function for $\vec{X}$
$P(\vec{X} \mid \omega_k)$	Probability density for $\vec{X}$ the class k. $\omega_k = E \in T_k$.
h(n)	A histogram of random values for the feature n.
$h_k(n)$	A histogram of random values for the feature n for the class k. $h(x) = \sum_{k=1}^{K} h_k(x)$
Q	Number of cells in h(*n*). $Q = N^D$

Bayesian Classification (Reminder)

Our problem is to build a box that maps a set of features $\vec{X}$ from an Observation, E into a class T_k from a set of K possible Classes.



Let ω_k be the proposition that the event belongs to class k: $\omega_k = E \in T_k$

ω_k Proposition that event E $\in$ the class k

In order to minimize the number of mistakes, we will maximize the probability that $\omega_k \equiv E \in T_k$

$$\hat{\omega}_k = \arg\max_k \left\{ \Pr(\omega_k \mid \vec{X}) \right\}$$

We will call on two tools for this:

1) Baye's Rule :

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

2) Normal Density Functions

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

Last week we looked at Baye's rule. Today we concentrate on Normal Density Functions.

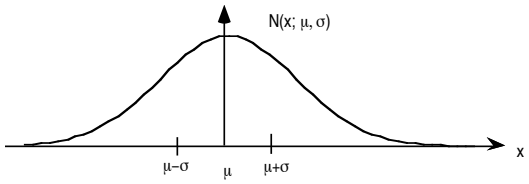
Normal Density Functions

Whenever a random variable is determined by a sequence of independent random events, the outcome will be a Normal or Gaussian density function. This is demonstrated by the Central Limit Theorem. The essence of the derivation is that repeated convolution of any finite density function will tend asymptotically to a Gaussian (or normal) function.

The exception is the dirac delta P(X) = δ(X).
In all other cases:

$$as\ N \to \infty \quad P(X)^{*N} \to \mathcal{N}(x;\ \mu,\ \sigma)$$

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



The parameters of $\mathcal{N}(x;\ \mu, \sigma)$ are the first and second moments.

Assume M observations $\{X_m\}$ for which we compute

The average value is the first moment of the samples

The "expected value" for $\{X_m\}$, E{X} is defined as the average or the mean:

$$E\{X\} = \frac{1}{M}\sum_{m=1}^{M} X_m$$

$\mu_x = E\{X\}$ is the first moment (or center of gravity) of $\{X_m\}$.

This is also true for a histogram. Map $\{X_m\} \rightarrow \{n_m\}$ in the range [1, N] as described above and compute the histogram h(n).

$$\forall m = 1, M : h(n_m) \leftarrow h(n_m) + 1$$

The mass of the histogram is the zeroth moment, M

$$M = \sum_{n=1}^{N} h(n)$$

The center of gravity (or mean or average) is the first moment μ_n

$$\mu_n = \frac{1}{N}\sum_{n=1}^{N} h(n) \cdot n$$

This is also the expected value of n.

$$\mu_n = E\{n\} = \frac{1}{M}\sum_{m=1}^{M} n_m$$

Thus the center of gravity of the histogram is the expected value of the random variable:

$$\mu_n = E\{n\} = \frac{1}{M}\sum_{m=1}^{M} n_m = \frac{1}{N}\sum_{n=1}^{N} h(n) \cdot n$$

And of course, the same is true for the continuous random variable $\{X_m\}$ and the pdf P(X).

$$E\{X\} = \frac{1}{M}\sum_{m=1}^{M} X_m = \int_{-\infty}^{\infty} P(X) \cdot X \, dX$$

Note that for a pdf the mass is 1 by definition: $S = \int_{-\infty}^{\infty} P(X)\, dX = 1$

The variance is the second moment of the samples

A similar relation exists for the Variance or Second Moment: σ.

For a set of observations of continuous random variable $\{X_m\}$
The variance is the "expected value" for of the squared difference from the average.

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

For a histogram h(n), from $\{X_m\} \rightarrow \{n_m\}$ in the range [1, N] as described above

The second moment is

$$\sigma_n^2 = \frac{1}{N}\sum_{n=1}^{N}h(n)\cdot(n - \mu_n)^2$$

This is also the variance of the set $\{n_m\}$ of samples.

$$\sigma_n^2 = E\{(n - \mu_n)^2\} = \frac{1}{M}\sum_{m=1}^{M}(n_m - \mu_n)^2$$

Thus the variance of the sample set is the second moment of the histogram

$$\sigma_n^2 = E\{(n - \mu_n)^2\} = \frac{1}{M}\sum_{m=1}^{M}(n_m - \mu_n)^2 = \frac{1}{M}\sum_{n=1}^{N}h(n)\cdot(n - \mu_n)^2$$

And of course, the same is true for the continuous random variable $\{X_m\}$ and the pdf P(X).

$$\sigma_x^2 = E\{(X - \mu_x)^2\} = \frac{1}{M}\sum_{m=1}^{M}(X_m - \mu_x)^2 = \int_{-\infty}^{\infty}P(X)\cdot(X - \mu_x)^2\,dX$$

Multi-Variate Normal Density Functions

In most practical cases, an observation is described by D features.
In this case a training set $\{\vec{X}_m\}$ an be used to calculate an average feature $\vec{\mu}$

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

If the features are mapped onto integers from [1,N]: $\{\vec{X}_m\} \rightarrow \{\vec{n}_m\}$ we can build a multi-dimensional histogram using a D dimensional table:

$$\forall m = 1,M : h(\vec{n}_m) \leftarrow h(\vec{n}_m) + 1$$

As before the average feature vector, $\vec{\mu}$, is the center of gravity (first moment) of the histogram.

$$\mu_d = E\{n_d\} = \frac{1}{M}\sum_{m=1}^{M} n_{dm} = \frac{1}{M}\sum_{n_1=1}^{N}\sum_{n_2=1}^{N}\dots\sum_{n_D=1}^{N} h(n_1,n_2,\dots,n_D)\cdot n_d = \frac{1}{M}\sum_{\vec{n}=1}^{N} h(\vec{n})\cdot n_d = \mu_d$$

$$\vec{\mu} = E\{\vec{n}\} = \frac{1}{M}\sum_{m=1}^{M}\vec{n}_m = \frac{1}{M}\sum_{\vec{n}=1}^{N} h(\vec{n})\cdot\vec{n} = \begin{pmatrix}\frac{1}{M}\sum_{\vec{n}=1}^{N} h(\vec{n})\cdot n_1\\ \frac{1}{M}\sum_{\vec{n}=1}^{N} h(\vec{n})\cdot n_2\\ \dots\\ \frac{1}{M}\sum_{\vec{n}=1}^{N} h(\vec{n})\cdot n_D\end{pmatrix} = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix}$$

For Real valued X:

$$\mu_d = E\{X_d\} = \frac{1}{M}\sum_{m=1}^{M} X_{dm} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} P(x_1,x_2,\dots x_D)\cdot x_d\, dx_1,dx_2,\dots,dx_D$$

In any case:

$$\vec{\mu} = E\{\vec{X}\} = \begin{pmatrix}E\{X_1\}\\ E\{X_2\}\\ \dots\\ E\{X_D\}\end{pmatrix} = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix}$$

For D dimensions, the second moment is a co-variance matrix composed of N^2 terms:

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

This is often written

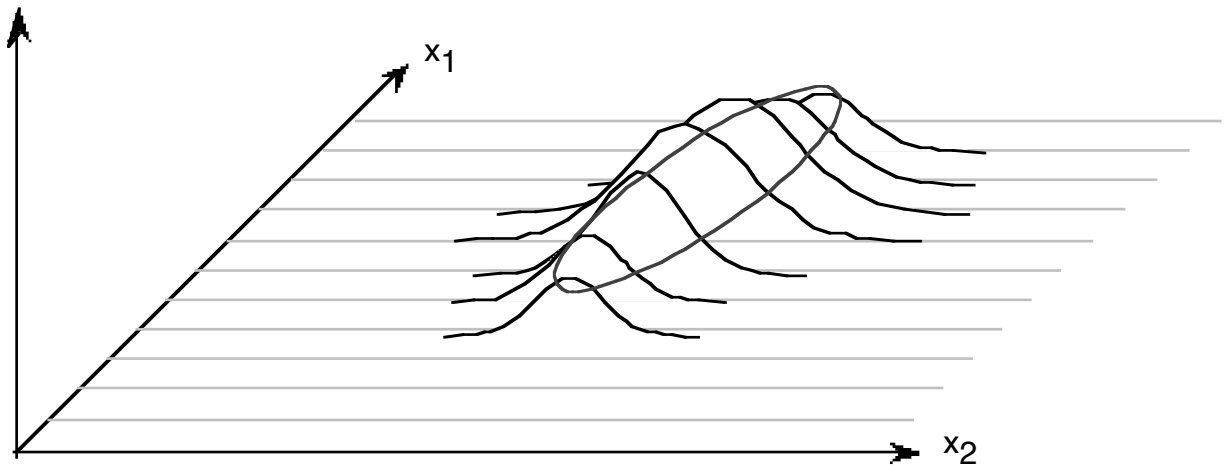
$$C_X = E\{(\vec{X} - E\{\vec{X}\})(\vec{X} - E\{\vec{X}\})^T\}$$

and gives

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

This provides the parameters for

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



The term $(2\pi)^{\frac{D}{2}} \det(\mathbf{C}_X)^{\frac{1}{2}}$ is a normalization factor.

$$\int\int \dots \int e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T C_X^{-1} (\vec{X}-\vec{\mu})} dx_1 dx_2 \dots dx_D = (2\pi)^{\frac{D}{2}} \det(C_X)^{\frac{1}{2}}$$

The determinant, det(C) is an operation that gives the volume of C.

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

for D=3

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

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

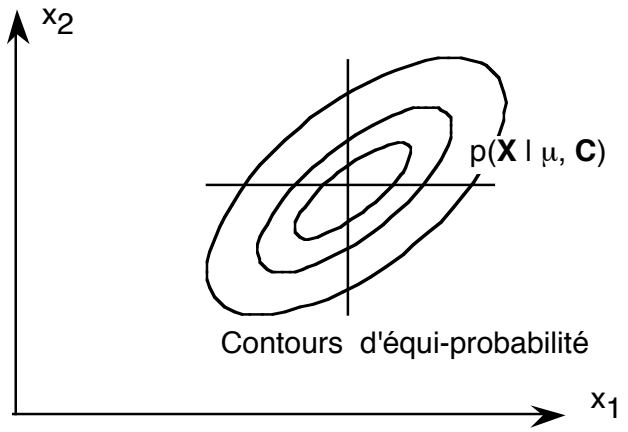
for D > 3 this continues recursively.

The exponent is positive and quadratic (2nd order). This value is known as the "Distance of Mahalanobis".

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

This is a distance normalized by the covariance. In this case, the covariance is said to provide the distance metric. This is very useful when the components of X have different units.

The result can be visualized by looking a equi-probably contours.



The matric C is positive and semi-definite.
The det(C) ≥ 0

If x_i and x_j are statistically independent, then $\sigma_{ij}^2 = 0$
For positive values of σ_{ij}^2, x_i and x_j vary together: For example Height and weight
For negative values of σ_{ij}^2, x_i and x_j vary in opposite directions.

Linear Algebraic Form for Moment Calculation

Calculation of the mean and covariance are often expressed using linear algebra. Such expressions are widely used in machine learning.

As before, assume a set of M_k training examples for the class k. $\{X_m^k\}$
The complete set of M examples is $\{X_m\} = \bigcup_{k=1,K} \{X_m^k\}$

Recall

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

Let us define $\vec{V}_m = \vec{X}_m - E\{\vec{X}_m\} = \vec{X}_m - \vec{\mu}_m$

and thus

$$C_x = E\{\vec{V}\vec{V}^T\}$$

We can compose a matrix with M columns and D rows from $\{V_m\}$.

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

This can be used to write

$$\mathbf{C_X} \equiv \mathbf{V}\ \mathbf{V}^{\mathrm{T}} = \begin{bmatrix}\bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet\end{bmatrix} \begin{bmatrix}\bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\end{bmatrix} = \begin{bmatrix}\bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\\ \bullet & \bullet & \bullet & \bullet\end{bmatrix}$$

$\mathbf{C_X} \equiv \mathrm{V}\ \mathrm{V}^{\mathrm{T}}$ is a D x D matrix that captures the "co-variance" of the elements of i,j of the vector X in $\{X_m\}$

Note that we can also write $\mathbf{C}_m = V^T V$ of size M x M.

C_m is the co-variance of the vectors in {X}.

Linear Transforms of the Normal Multivariate Density

The Normal (Gaussian) function is a defined only by its moments.
It is thus invariant to transformations of its moments, that is affine transformations.
The affine transformations include rotation, translation, scale changes and other linear transformations.

For example consider a rotation vector of cosine angles about each component of X :

$$\vec{R} = \begin{pmatrix} \cos(\alpha_1) \\ \cos(\alpha_2) \\ ... \\ \cos(\alpha_D) \end{pmatrix}$$

Note that $\| \vec{R} \| = 1$.

Rotation is projection onto $\vec{R}$

A vector $\vec{X}$ may be rotated by $\vec{R}$

$$\vec{Y} = \vec{R}^T \vec{X}$$

For the covariance: $C_Y = E\{(\vec{R}^T\vec{V})(\vec{R}^T\vec{V})^T\}$

$$= E\{(\vec{R}^T\vec{V})(\vec{V}^T\vec{R})\}$$
$$= E\{(\vec{R}^T(\vec{V}\vec{V}^T)\vec{R})\}$$
$$= E\{(\vec{R}^T C_X \vec{R})\}$$

$(\vec{R}^T \vec{V})^T = (\vec{V}^T \vec{R})$

Thus rotation of a covariance requires pre and post multiplication by $\vec{R}$.

Projection of a Gaussian is the Gaussian of the projection.

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

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

This can be computed by projecting the moments or by projecting the data and recomputing the moments.

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