Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2010/2011

Lesson 16 15 april 2011

Normal Probability Density Functions

Sources Bibliographiques :

"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.

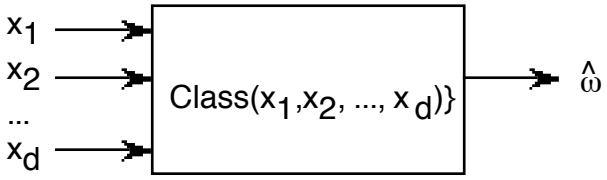
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

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.
C_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 C_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

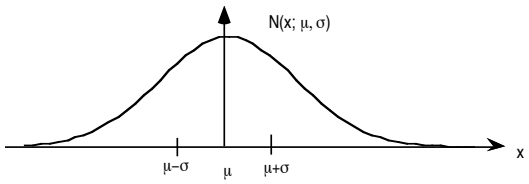
$$P(\vec{X} \mid \omega_k) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(C_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T C_k^{-1} (\vec{X}-\vec{\mu}_k)}$$

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

The Normal (Gaussian) Density Function

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.

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



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

This is often written as a conditional:

This is sometimes expressed as a conditional $\mathcal{N}(X \mid \mu,\, \sigma)$

In most cases, for any density p(X) :

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

This is the Central Limit theorem.

Multivariate 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 D^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

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

and gives

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

This provides the parameters for

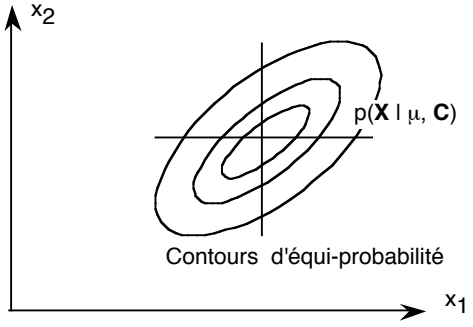
$$p(\vec{X}) = \mathcal{N}(\vec{X} \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1} (\vec{X}-\vec{\mu})}$$

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

$$d(\vec{X}; \vec{\mu}, C)^2 = -\frac{1}{2}(\vec{X} - \vec{\mu})^T \Sigma^{-1} (\vec{X} - \vec{\mu})$$

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 at the equi-probably contours.



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 negative values of σ_{ij}^2, x_i and x_j vary in opposite directions.

For example, consider features x_1 = height *(m)* and x_2 = weight *(kg)*

In most people height and weight vary together and so σ_{12}^2 would be positive

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 (cosine) vector $\vec{R}$, such that $\| \vec{R} \| = 1$.

Rotation is projection onto a vector $\vec{R}$

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

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

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

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 :

Projection of a Gaussian is the Gaussian of the projection.

For a projection R : $\vec{Y} = \vec{R}^T \vec{X}$

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

Note that for the Covariance, projection requires pre- and post- multiplication by $\vec{R}$.
We can demonstrate this with a linear algebraic expression of the moments.

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 training examples $\{X_m\}$

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

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

$$X = \begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1M}\\ x_{21} & x_{22} & \cdots & x_{2M}\\ \cdots & \cdots & & \cdots\\ x_{D1} & x_{D2} & \cdots & x_{DM}\end{pmatrix}$$ Let us define the unity vector : $\vec{u} = \begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix}$

Then $$\vec{\mu} = \begin{pmatrix}\mu_1\\ \mu_2\\ \dots\\ \mu_D\end{pmatrix} = \begin{pmatrix}x_{11} & x_{12} & \cdots & x_{1M}\\ x_{21} & x_{22} & \cdots & x_{2M}\\ \cdots & \cdots & & \cdots\\ x_{D1} & x_{D2} & \cdots & x_{DM}\end{pmatrix}\begin{pmatrix}1\\ 1\\ \vdots\\ 1\end{pmatrix} = X\cdot\vec{u}$$

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

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

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

From this: $$\Sigma_x = E\{\vec{V}\vec{V}^T\}$$

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

This can be seen as

$$\Sigma_x \equiv V\ V^T = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{bmatrix} = \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet \end{bmatrix}$$

Note that we can also write $\Sigma_m = V^T\ V$ of size M x M.

We can use this to show that projection of a covariance requires pre and post multiplication:

$$\sigma_y^2 = \vec{R}^T \Sigma_X \vec{R}$$

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

For the covariance:
$$\begin{aligned} \Sigma_Y &= E\{(\vec{R}^T\vec{V})(\vec{R}^T\vec{V})^T\} \\ &= E\{(\vec{R}^T\vec{V})(\vec{V}^T\vec{R})\} \\ &= E\{(\vec{R}^T(\vec{V}\vec{V}^T)\vec{R})\} \\ &= E\{(\vec{R}^T\Sigma_X\vec{R})\} \end{aligned}$$

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