Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2013/2014

Lesson 15 9 april 2014

Normal (Gaussian) 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 k
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that $E \in C_k$
$P(\omega_k) = P(E \in C_k)$	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. $\omega_k = E \in C_k$.
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*). $Q = N^D$

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

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

Probability Density Functions

The alternative to a non-parametric representation is to use a function to represent $p(\vec{X} \mid \omega_k)$ and $p(\vec{X})$. Such a function is referred to as a "Probability Density Function" or PDF. Note that we use upper case for probabilities and lower case for functions. *P(ω)* is a value. *p(X)* is a function.

A probability density function of a continuous random variable is a function that describes the relative likelihood for this random variable to occur at a given point in the observation space. "Likelihood" is a relative measure of belief or certainty.

Note that Likelihood is not probability. We will use the "likelihood" to determine the parameters for parametric models of probability density functions. To do this, we first need to define probability density functions.

A probability density function, $p(\vec{X})$, is a function of a continuous variable or vector, $\vec{X} \in R^D$, of random variables such that :

1) $\vec{X}$ is a vector of D real valued random variables with values between $[-\infty, \infty]$
2) $\int_{-\infty}^{\infty} p(\vec{X}) = 1$

Note that, $p(\vec{X})$ is NOT a number but a continuous function. To obtain a probability we must integrate over some volume V of the D dimensional feature space.

$$P(\vec{X} \in V) = \int_V p(\vec{X}) d\vec{X}$$

This integral gives a number that can be used as a probability.

In the case of D=1, the probability that X is within the interval [A, B] is

$$P(X \in [A,B]) = \int_A^B p(x) dx$$

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

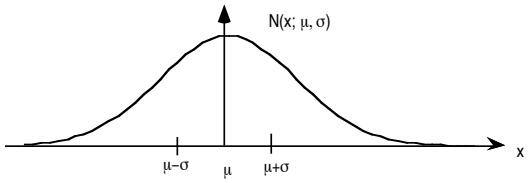
While $p(\vec{X})$ and $p(\vec{X} \mid \omega_k)$ are NOT numbers, their ratio IS a number.

The ratio of two pdf's gives a probability value!

The Univariate Normal 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.

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



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

According to the Central Limit theorem., for any real density p(X) :

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

$\mathcal{N}(x;\ \mu,\ \sigma)$ is composed of 3 parts:

1) e 2) $\frac{1}{\sqrt{2\pi}\sigma}$ 3) $-\frac{(X-\mu)^2}{2\sigma^2}$

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

1) Eulers Number "e"

e = 2.7818281828…

e is an irrational and transcendental constant approximately equal to 2.718281828….

Sometimes referred to as Euler's Number, e has many useful properties.
For use in the Normal density, e simplifies the algebra.

$$\int_{-\infty}^{\infty} e^x dx = e^x$$

2) a normalization factor:

$\frac{1}{\sqrt{2\pi}\sigma}$ is a normalization factor

$$\sqrt{2\pi}\sigma = \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

3) The Mahalanobis distance. This is where the action is.

$$d(x,\mu;\sigma)^2 = \frac{(x-\mu)^2}{2\sigma^2}$$

This is the difference between x and μ normalized by the σ.

The normal density is e to the negative power of a distance.

To better understand, we need to talk about expected values, and moments.

Expected Values and Moments

The average value is the first moment of the samples

For M samples of a numerical feature value $\{X_m\}$, the "expected value" 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 the values of $\{X_m\}$.
$\mu_x = E\{X\}$ is also the first moment (or center of gravity) of the resulting pdf.

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

This is also true for histograms.
Suppose that the feature, X, is an integer with values from [1, N].

We build the histogram as before: $\forall m=1, M : h(x_M) := h(x_M) + 1;$
The mass of the histogram is M

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

M is also the number of samples used to compute h(n).
The expected value of $\{X_m\}$ is the 1st moment of h(n).

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

The same is true for the variance
The variance is the expected square of the deviation from the average

$$\sigma^2 = E\{(X-\mu)^2\} = E\{X^2\} - \mu^2 = E\{X^2\} - E\{X\}^2$$

This is also the second moment of the pdf

$$\sigma^2 = E\{(X-\mu)^2\} = \frac{1}{M}\sum_{m=1}^{M}(X_m-\mu)^2 = \int_{-\infty}^{\infty} p(x)\cdot(x-\mu\ )^2\ dx$$

and the second moment of a histogram for discrete features.

$$\sigma^2 = E\{(X-E\{X\})^2\} = \frac{1}{M}\sum_{m=1}^{M}(X_m-\mu)^2 = \frac{1}{M}\sum_{n=1}^{N} h(n)\cdot(n-\mu)^2$$

Biased and Unbiased Variance

Note that this is a "Biased" variance. The unbiased variance would be

$$\tilde{\sigma}^2 = \frac{1}{M-1}\sum_{m=1}^{M}(X_m - \mu)^2$$

If we draw a random sample $\{X_m\}$ of M random variables from a Normal density with parameters (μ, σ)

$$\{X_m\} \leftarrow \mathcal{N}(x;\mu,\tilde{\sigma})$$

Then we compute the moments, we obtain.

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

and

$$\hat{\sigma}^2 = \frac{1}{M}\sum_{m=1}^{M}(X_m - \mu)^2 \qquad \text{Where } \tilde{\sigma}^2 = \frac{M}{M-1}\hat{\sigma}^2$$

Note the notation: ~ means "true", ^ means estimated.

The expectation underestimates the variance by 1/M.

The RMS error for estimating $p(X)$ from M samples $\{X_m\}$ is the difference between a biased and unbiased error.

Multivariate Normal Density Functions

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

as with univariate, there are 3 parts to $\mathcal{N}(\vec{X};\vec{\mu},\,\Sigma)$:

$$\frac{1}{(2\pi)^{\frac{D}{2}}\det(\Sigma)^{\frac{1}{2}}}, \quad e, \quad \text{and} \quad d(\vec{X};\vec{\mu},\Sigma)^2 = -\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1}(\vec{X}-\vec{\mu})$$

1) e = 2.7818281828…

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

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

The determinant, det($\boldsymbol{\Sigma}$) is an operation that gives the volume of $\boldsymbol{\Sigma}$.

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.

3) The Mahalanobis distance. This is where the action is.

The Multivariate Mahalanobis Distance

The exponent of $\mathcal{N}(\vec{X};\vec{\mu}, \Sigma)$ is known as the "Distance of Mahalanobis".

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

This is a distance normalized by the covariance. It is positive and quadratic (2nd order).

The covariance is said to provide the distance metric. This is very useful when the components of X have different units.

When an observation is described by D features, the training set $\{\vec{X}_m\}$ an be used to calculate an average feature $\vec{\mu}$

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

$\vec{\mu}$, is the center of gravity (first moment) of the pdf. This is the vector of averages for the components, X_d of $\vec{X}$

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

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.

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

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

In either case: $$\vec{\mu} = E\{\vec{X}\} = \begin{pmatrix} E\{X_1\} \\ E\{X_2\} \\ \cdots \\ E\{X_D\} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \cdots \\ \mu_D \end{pmatrix}$$

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

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

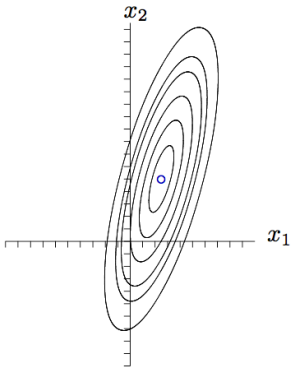
and gives

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

This provides the parameters for

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

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



Ellipses for 99%, 95%, 90%, 75%, 50%, and 20% of the mass

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

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

A vector $\vec{X}$ may be projected into a space $\vec{Y}$ by

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

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

For a projection onto a 1D vector Y, R is D x 1 : $\vec{Y} = \vec{R}^T \vec{X}$

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

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 $\{X_m\}$.

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

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

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 $\{V_m\}$.

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

From this: $\Sigma_x = E\{\vec{V}\vec{V}^T\}$ can be computed as a vector product.

$\Sigma_x \equiv V\ V^T$ 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

$$\Sigma_X = \mathbf{V}\mathbf{V}^T = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

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:

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

Thus
$$\Sigma_y = (\vec{R}^T V)(\vec{R}^T V)^T$$
$$\Sigma_y = (\vec{R}^T V)(V^T \vec{R})$$
$$\Sigma_y = \vec{R}^T (VV^T)\vec{R}$$
$$\Sigma_y = \vec{R}^T \Sigma_X \vec{R}$$

Thus projection of a covariance requires pre and post multiplication by $\vec{R}$.
In the case of projection to a 1D vector Y:

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