Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2011/2012

Lesson 15 4 april 2012

Expect Values and 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

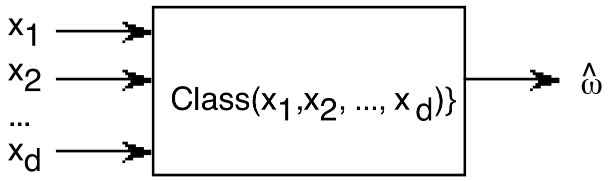
x	a variable
X	a random variable (unpredictable value)
N	The number of possible values for X (Can be infinite).
$\vec{x}$	A vector of D variables.
$\vec{X}$	A vector of D random variables.
D	The number of dimensions for the vector $\vec{x}$ or $\vec{X}$
E	An observation. An event.
C_k	The class k
k	Class 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 $h(n)$. $Q = N^D$

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

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

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 the event E $\in$ the class k

In order to minimize the number of mistakes, we will maximize the probability that that the event E $\in$ the class k

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

We will rely 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

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

Today we concentrate on Normal Density Functions.

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 $\{X_m\}$.

This can be seen from the histogram h(x).
The mass of the histogram is the zeroth moment, M

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

M is also the number of samples used to compute h(n).
The expected value of is the average μ

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

This is also the expected value of n.

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

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

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

The second moment is the expected deviation from the first moment:

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

Expected Values for PDFs

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) $\int_{-\infty}^{\infty} p(\vec{X}) = 1$

We cab replace $\frac{1}{M} h(\vec{X}) \rightarrow p(\vec{X})$

For Bayesian conditional density $\frac{1}{M_k} h(\vec{X} \mid \omega_k) \rightarrow p(\vec{X} \mid \omega_k)$

Where $p(\vec{X})$, has a mass of 1 by definition:

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

Just as with histograms, the expected value is the first moment of a pdf.

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

The second moment is

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

Note that

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

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

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

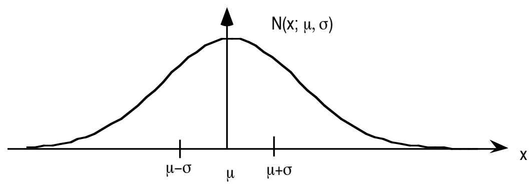
Note the notation: ~ means "true", ^ means estimated.

The expectation underestimates the variance by 1/M.
Later we will see that the expected RMS error for estimating p(X) from M samples is related to the bias. But first, we need to examine the Gaussian (or normal) density function.

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.

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



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) :

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

This is the Central Limit theorem.
An exception is the dirac delta $p(X) = \delta(x)$.

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 $\vec{\mu}$

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

If the features are mapped onto integers from [1, N]: $\{\vec{X}_m\} \to \{\vec{n}_m\}$ 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.

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

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

For Real valued X:

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

In any case:

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

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

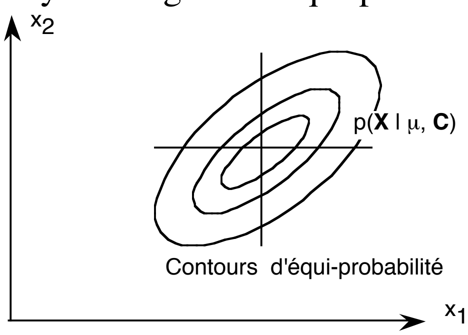
$$p(\vec{X}) = \mathcal{N}(\vec{X} \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1} (\vec{X}-\vec{\mu})}$$

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