Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2009/2010

Lesson 14 2 april 2010

Baye's Rule with Normal Density Functions

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*). $Q = N^D$

Bayes Rule as a Ratio of Histograms

Histograms provide an alternate view of Baye's Rule.
This view illustrates how Baye's rule can be used with density functions (pdf's)

Histograms

As we saw, for integer x from a bounded set of values, such that $x \in [1, N]$:
Given a training set $\{X_m\}$ of features from M events, we can build a table of frequency for the values of X.

For m=1 to M $h(X_m) = h(X_m)+1;$

the probability that a feature $X \in \{X_m\}$ from this set has the value x is then

$$P(X=x) = \frac{1}{M} h(x)$$

If the
1) the sample is large enough ($M > 10\ Q$, where $Q=N^D$), and
2) the observing conditions are "ergodic" (do not change with time),
then
the histogram will also predict frequency of occurrence for future events.

The validity of this depends on the ratio of the number of sample observations M and the number of cells in the histogram $Q=N^D$. This is true for vectors (D>1) as well as scalar features (D=1).

For a vector of D values $\vec{x}$ the table has D dimensions. $h(x_1, x_2, \ldots, x_D) = h(\vec{x})$

The average error depends on the ration $Q=N^D$ and M. : $E_{ms} \sim O(\frac{Q}{M})$

We need to assure that $M >> Q = N^D$

A general rule is $M \geq 10N^D$
In many examples, we will prefer to propose $M \geq 8 \cdot N^D$ because $8 = 2^3$ will be easier to manipulate when using exponential representations for very large numbers (see below).

Example:

Suppose that we have 2 classes, k=1 and k=2, and that we observe
M_1 events from class k=1: $\{X_m^1\}$ and M_2 events from class k=2 $\{X_m^2\}$
Assume ergodic observing conditions with $M_1 \geq 8N$ and $M_2 \geq 8N$

We maps the features X to integers : $\{n_m^1\}$ and $\{n_m^2\}$ in the range [1, N].

We build the histograms $h_1(n)$ and $h_2(n)$:

for m=1 to M_1 : $h_1(n_m^1) := h_1(n_m^1) + 1$
for m=1 to M_2 : $h_2(n_m^2) := h_2(n_m^2) + 1$

We also define $h(n) = h_1(n) + h_2(n)$ and $M = M_1 + M_2$

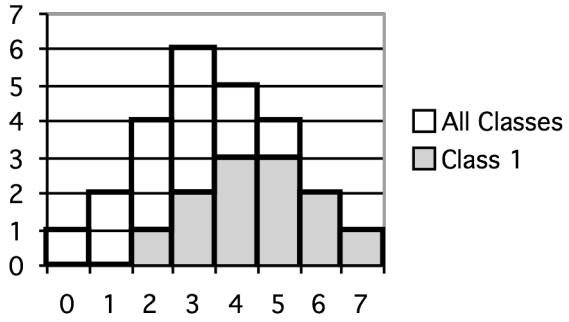
We note that the $p(E \in T_1) = p(\omega_1) = \frac{M_1}{M}$ and $p(E \in T_2) = p(\omega_2) = \frac{M_2}{M}$

Thus, for a new observation, E, with feature X mapped to "n", then

$$p(n) = \frac{1}{M} h(n)$$
$$p(n \mid \omega_1) = \frac{1}{M_1} h_1(n)$$

Thus

$$p(\omega_1 \mid n) = \frac{p(n \mid \omega_1) p(\omega_1)}{p(n)} = \frac{\frac{1}{M_1} h_1(n) \frac{M_1}{M}}{\frac{1}{M} h(n)} = \frac{h_1(n)}{h(n)}$$



For example, $p(\omega_1 \mid n=2) = ¼$
The probability of observing class k give feature n is $p(w_k|n) = h_k(n)/h(n)$

Multi-dimensional histograms

This method can be generalized to vectors with any number of dimensions.

ATTENTION! The number of cells will grow exponentially with D.

The histogram must have sufficient samples M.

$$M \geq 8\,Q = 8\,N^{D.}$$

Here is a table of numbers of cells, Q, in a histogram of D dimensions of N values.

N \ D	D=1	D=2	D=3	D=4	D=5	D=6
N=2	2^1	2^2	2^3	2^4	2^5	2^6
N=4	2^2	2^4	2^6	2^8	2^{10} =1 Kilo	2^{12} =2 Kilo
N=8	2^3	2^6	2^9	2^{12}	2^{15}	2^{18}
N=16	2^4	2^8	2^{12}	2^{16}	2^{20} = 1 Meg	2^{24} = 4 Meg
N=32	2^5	2^{10} =1 Kilo	2^{15}	2^{20} = 1 Meg	2^{25}	2^{30} = 1 Gig
N=64	2^6	2^{12}	2^{18}	2^{24}	2^{30} = 1 Gig	2^{36}
N=128	2^7	2^{14}	2^{21} = 2 Meg	2^{28}	2^{35}	2^{42} =2 Tera
N=256	2^8	2^{16}	2^{24}	2^{32} = 2 Gig	2^{40} = 1 Tera	2^{48}

For a problem with D dimensions, and M samples, choose N such a that:

$$M \geq 8N^D \Rightarrow N = \mathrm{Log}_2(M) - 3$$

Images contain a LOT of data (≥ 1 meg pixel/image). Thus histograms have been used for many years for image analysis and computer vision.

As computing power and memory have grown, histograms have emerged as ideally suited for use with very large scale data sets as provided by the World Wide Web and social networks.

Bayes Rule with Density Functions.

There are problems for which it is not possible to map the features to a finite set of integer values. There are also problems for which sufficient data is not available. There are also cases for which the observation is NOT ergodic.

What can we do when $M \leq 8N^D$?

We can generalize the h(n) as P(X), a probability density function (pdf):
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) $\int_{-\infty}^{\infty} P(\vec{X}) = 1$

In this case we replace $\frac{1}{M} h(n) \rightarrow P(\vec{X})$ and $\frac{1}{M_k} h(n \mid \omega_k) \rightarrow P(\vec{X} \mid \omega_k)$

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

Note that the ratio of two pdfs gives a probability value!

This will be our primary tool for designing recognition machines.

There is one more tool we need : Normal density Functions:

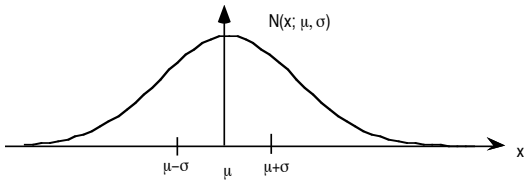
Normal Density Functions

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).

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

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

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

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