Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2009/2010

Lesson 13 31 March 2010

Two Views on Bayes Rule

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.
A,B	Classes of events.
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 $\vec{X}$ given ω_k where $\omega_k = E \in T_k$.
h(x)	A Histogram of values for x.
$h_k(\vec{X})$	A histogram of values for x for the class k. $h(x) = \sum_{k=1}^{K} h_k(x)$
Q	Number of cells in $h(\vec{X})$. $Q = N^D$

Probability

A probability is a form of truth function.
A probability maps a proposition into a numerical truth value between 0 and 1.

Formally, the proposition should be a predicate, such as
Member(E, Class-K) or Equals(X=y).

For notational reasons, this is often simplified.
For example, for class membership, we will use the greek variable ω_k
for Member(E, Class-K).

Thus $p(\omega_k) = p(E \in T_k) = p(\text{Member}(E, \text{Class-K}))$

Probability will be represented by a lower case function p().

There are two ways to define the meaning of "probability":

1) Statistics : Frequency of Occurrence: The fraction of times that something is true, or
2) Probability axioms: Using a systems of axioms.

A statistical, or Frequency-based, approach is more intuitive, but not always possible to apply. In some cases, the axioms of probability theory can provide a solution that is not possible from frequency of occurrence.

In either case, probability is a function that returns a number between 0 and 1.
$\Pr() \in [0, 1]$.

Probability as Frequency of Occurence.

A frequency based definition of probability is sufficient for many practical problems.

Suppose we have M observations of random events, $\{E_m\}$, for which M_k of these events belong to the class k. The probability that one of these observed events belongs to the class k is:

$$\Pr(E \in T_k) = \frac{M_k}{M}$$

If we make new observations under the same observations conditions (ergodicity), then it is reasonable to expect the fraction to be the same. However, because the observations are random, there may be differences. These differences will grow smaller as M grows larger.

The average (root-mean-square) error for

$$\Pr(E \in T_k) = \frac{M_k}{M}$$

will be proportional to k and inversely proportional to M.

Axiomatic Definition of probability

Probability theory provides a more abstract definition for probability using 3 axioms.

An axiomatic definition makes it possible to apply analytical techniques to the design of classification systems. Only three postulates (or axioms) are necessary:
In the following, let E be an event, let S be the set of all events, and let T_k be set of events that belong to class k with K total classes. $S = \bigcup_{k=1,K} T_k$

Postulate 1 : $\forall\ T_k \in S\ :\ p(E \in T_k) \geq 0$
Postulate 2 : $p(E \in S) = 1$
Postulate 3 :
$\forall\ T_i, T_j \in S$ such that $T_i \cap T_j = \varnothing$: $p(E \in T_i \cup T_j) = p(E \in T_i) + p(E \in T_j)$

A probability function is any function that respect these three axioms.
A probability is the truth value produced by a probability function.

The probability of the value for a random variable.

Probability Density Function (pdf)

A probability density function (pdf) is a function of a continuous variable or vector, $\vec{X}$, of variables such that :

1) $\vec{X} \in R^D$: $\vec{X}$ is a vector of D real valued variables with values between $[-\infty, \infty]$

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

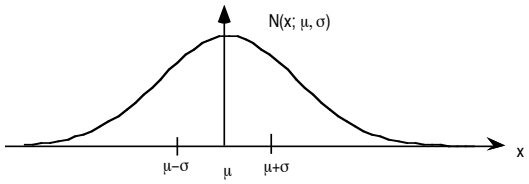
We will always note a pdf as a Capital P().

The simplest example is the uniform density function.

rect : P(X) = rect (X).

An important pdf is the Normal or Gaussian density function:

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



In a few weeks we will also use a Sum or Mixture of Normal density functions.

$$P(X) = \sum_{n=1}^{N} \alpha_n \mathcal{N}(x; \mu_n, \sigma_n)$$

Note that the ratio of two pdfs can be a probability:

$$p(E) = \frac{P1(X)}{P2(X)}$$

We will use pdf's to derive recognition systems. However, as an introduction, much of the intuition for pdf's can be gained from a simplification: histograms.

Histogram Representation of Probability of a variable

For integer x, such that x has a finite number of values $x \in [x_{min}, x_{max}]$, we can consider each value of x as a class.

We can then estimate the probability for each class using M observations $\{X_m\}$.

To estimate the probability of a value, we count the number of times it occurs.
For this we use a table of "frequency of occurrence", also known as a "histogram", h(x).

The existence of computers with gigabytes of memory has made the computation of such tables practical.

We use the table to count the number of times each value occurs:

$$\forall m=1, M \; : \; h(X_m) := h(X_m) + 1; \; M := M+1;$$

Thus, the probability of a value, $X \in [X_{min}, X_{max}]$ is the frequency of occurrence of the value.

The probability that a random value X takes a given value x is

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

Problem: How many observations, M, do we need?

Answer:

Given N possible values of x, and M observations, in the worst case:

"average error" is proportional to O(N/M).

Rule of thumb. For most applications, we need M = 10N (10 samples per "cell").

The Classic Formulation for Bayes Rule

Let E represent a random event from an event "generator" (for example a sensor).

Consider 2 independent classes of events A and B such that

E may be $A \cap B$ or $\neg A \cap B$ or $A \cap \neg B$ or $\neg A \cap \neg B$

For these two classes we can form 2 propositions, q and r.

$q \equiv E \in A$ and $r \equiv E \in B$

thus the probability of each proposition is

$p(q) \equiv Pr\{E \in A\}$ and
$p(r) \equiv Pr\{E \in B\}$.

From the axioms of probability:

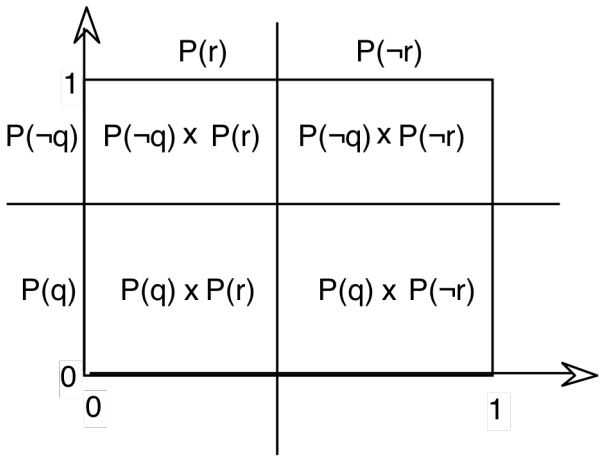
$p(q) + p(\neg q) = 1$.

$p(q \wedge r)$ is the joint probability of q and r, that $E \in A$ and $E \in B$

If A and B are independent then q and r are independent :

$p(q \wedge r) = p(q) \cdot p(r)$,
$p(q \vee r) = p(q) + p(r)$.

Graphically this is drawn:



$$p(q \wedge r) + p(q \wedge \neg r) + p(\neg q \wedge r) + p(\neg q \wedge \neg r) = 1$$

The marginal probabilities are:

$$p(q) = p(q \wedge r) + p(q \wedge \neg r)$$
$$p(r) = p(q \wedge r) + p(\neg q \wedge r)$$

The "conditional" probabilities are defined as :

$$p(r \mid q) = \frac{p(q \wedge r)}{p(q)} = \frac{p(q \wedge r)}{p(q \wedge r) + p(q \wedge \neg r)}$$

and

$$p(q \mid r) = \frac{p(q \wedge r)}{p(r)} = \frac{p(q \wedge r)}{p(q \wedge r) + p(\neg q \wedge r)}$$

that is, the probability that q is true, given that r is true is p(q | r).

By algebra :

$$p(r \mid q)\, p(q) = p(q \wedge r) = p(q \mid r)\, p(r)$$

thus

$$p(r \mid q)\, p(q) = p(q \mid r)\, p(r)$$

This is Bayes Rule. This can also be written:

$$p(r \mid q) = \frac{p(q \mid r) p(r)}{p(q)}$$

Bayes Rule as a Ratio of Histograms

Histograms provide an alternate view of Bayes Rule.
This view illustrates how Bayes rule can be used with pdfs

Histograms

As we saw, for integer x from a bounded set of values, such that $x \in [x_{min}, x_{max}]$,

the probability that a random observation X takes on x is

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

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

This is true for vectors as well as values.

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 $M >> Q = N^d$

As a general rule : $M = 10N^d$

Example:

Suppose that we have a bounded random variable X such that

$$0 \le X < x_{max}$$

We can map X to an integer n with N values between 0 and N–1 using trunc()

$$n = trunc(N \frac{X}{x_{max}})$$

Suppors 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 $\{X_m^2\}$

We maps these to integers : $\{n_m^1\}$ and $\{n_m^2\}$

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 $M = M_1 + M_2$

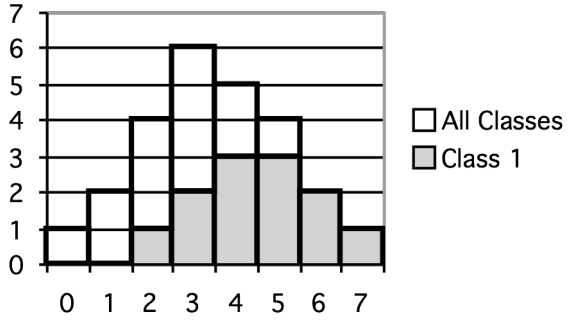
Note that the $p(E \in T_1) = p(\omega_1) = \frac{M_1}{M}$

Thus, for a new observation X, $n = trunc(N \frac{X}{x_{\max}})$ then

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

Thus the

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



For example, $p(\omega_1 \mid n=2) = 1/4$

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.

$M \geq 10 \qquad Q \geq 10\ N^{D.}$

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

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