Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2010/2011

Lesson 13 1 April 2011

Introduction to Bayesian Recognition

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 $\in T_k$
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. $\{X_m\} = \bigcup_{k=1,K} \{X_m^k\}$

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 M_k and inversely proportional to M.

Axiomatic Definition of probability

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.

Histogram Representation of Probability

We can use histograms both as a practical solution to many problems and to illustrate fundamental laws of axiomatic probability.

When we have K classes of events, we can build a table of frequency of occurrence for events from each class $h(E \in T_k)$.

The table of "frequency of occurrence" is also known as a "histogram", h(x).
The existence of computers with gigabytes of memory has made the computation of such tables practical.

The table h() can be implemented as a hash table, using the labels for each class as a key. Alternatively, we can map each class onto K natural numbers k <- T_k

$\forall m=1, M$: if $E_m \in T_k$ then h(k) := h(k) + 1;

After M events, given a new event, E,

$$p(E \in T_k) = p(k) = \frac{1}{M} h(k)$$

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

Answer: Given N possible values of X, h(x) has Q = N cells.

For M observations, in the worst case the RMS error between an estimated h(X) and the true h(x) is proportional to O(Q/M).

For most applications, $M \geq 10\ Q$ (10 samples per "cell") is reasonable.

Bayesian Probability

Bayesian probability can be seen as an extension of logic that enables reasoning with uncertain statements. Bayesian probability interprets probability as "a measure of a state of knowledge", rather than as "frequency of occurrence".

In Bayesian probability, the confidence of a proposition is represented by a probability number between 0 and 1.

To evaluate the confidence of a hypothesis, we determine a prior probability
This prior is then updated by observing new evidence.

The Bayesian interpretation provides a standard set of procedures and formulae to perform this calculation.

Although Bayesian logic is based on axiomatic probability, we can use histograms to illustrate Bayes rule.

Illustrating Bayes Rule with Histograms

Suppose we have a set of events described by a pair of properties.
For example, consider the your grade in 2 classes x_1 and x_2.

Assume your grade is a letter grade from the set {A, B, C, D, F}.

We can build a 2 dimensional hash table, where each letter grade acts as a key into the table $h(x_1, x_2)$.

This hash table has Q= 5 x 5 = 25 cells.

Each student is an observation with a pair of grades (x_1, x_2).

$$\forall m=1, M \text{ : if } h(x_1, x_2) := h(x_1, x_2) + 1;$$

Question: How many students are needed to fill this table?
Answer $M \geq 10Q = 250$.

An example, consider the table as follows:

		x_1					
	$h(x_1,x_2)$	A	B	C	D	F	$r(x_2)$
x_2	A	2	5	3	1		11
	B	5	16	8	1		30
	C	2	12	20	3	1	38
	D		2	6	2	2	12
	F			4	4	1	9
	$c(x_1)$	9	35	41	11	4	100

Any cell, (x_1, x_2) represents the probability that a student got grade X_1 for course C_1 and grade X_2 for course C_2.

$$p(X_1 = x_1 \wedge X_2 = x_2) = \frac{1}{M} h(x_1,x_2)$$

Let us note the sum of column x_1 as $c(x_1)$ and sum of row x_2 as $r(x_2)$ and the value of cell x_1,x_2 as $h(x_1,x_2)$

$$c(x_1) = \sum_{x_2=\{A,B,...F\}} h(x_1,x_2) \qquad r(x_2) = \sum_{x_1=\{A,B,...F\}} h(x_1,x_2)$$

for example $r(x_1=B) = 30$, $C(x_2=B) = 35$, $h(x_1,x_2) = 16$

From this table we can easily see three fundamental laws of probability:

Sum Rule: $$p(X_1 = x_1) = \sum_{x_2=\{A,B,...,F\}} p(X_1 = x_1, X_2 = x_2) = \frac{1}{M} \sum_{x_2=\{A,B,...,F\}} h(x_1,x_2) = \frac{1}{M} c(x_1)$$

example: $$p(x_1 = B) = \sum_{x_2=A,B,...,F} p(x_1 = B, x_2) = \frac{1}{M} \sum_{x_2=A,B,...,F} h(B,x_2) = \frac{c(B)}{M} = \frac{35}{100}$$

from which we derive the sum rule:

$$p(X_1 = x_1) = \sum_{X_2} p(X_1 = x_1, X_2 = x_2)$$

or more simply

$$p(X_1) = \sum_{X_2} p(X_1, X_2)$$

This is sometimes called the "marginal" probability, obtained by "summing out" the other probabilities.

Conditional probability:
We can define a "conditional" probability as the fraction of one probability given another.

$$p(X_1 = x_1 \mid X_2 = x_2) = \frac{h(x_1,x_2)}{r(x_2)} = \frac{h(x_1,x_2)}{\sum_{x_1} h(x_1,x_2)}$$

For example.

$$p(X_1 = B \mid X_2 = C) = \frac{h(B,C)}{\sum_{x_1} h(x_1,C)} = \frac{12}{38} \text{ and } p(X_2 = C \mid X_1 = B) = \frac{h(B,C)}{\sum_{x_2} h(B,x_2)} = \frac{12}{35}$$

From this, we can derive Bayes rule :

$$p(X_1 \mid X_2)\cdot p(X_2) = \frac{h(X_1,X_2)}{\sum_{X_1} h(X_1,X_2)} \cdot \sum_{X_1} h(X_1,X_2) = h(X_1,X_2) = \frac{h(X_1,X_2)}{\sum_{X_2} h(X_1,X_2)} \cdot \sum_{X_2} h(X_1,X_2) = p(X_2 \mid X_1)\cdot p(X_1)$$

or more simply

$$p(X_1 \mid X_2)\cdot p(X_2) = p(X_2 \mid X_1)\cdot p(X_1)$$

or more commonly written:

$$p(X_1 \mid X_2) = \frac{p(X_2 \mid X_1)\cdot p(X_1)}{p(X_2)}$$

Product Rule:

We can also use the histogram to derive the product rule.

Note that $p(X_1 = i, X_2 = j) = h(i,j)$

$$p(X_1 = i \mid X_2 = j) = \frac{h(i,j)}{\sum_i h(i,j)}$$

and

$$p(X_1,X_2) = p(X_1 \mid X_2)\cdot p(X_2)$$

These rules show up frequently in machine learning and Bayesian estimation.

Note that we did not need to use numerical values for x_1 or x_2.

If the features are symbolic, $h(x_1, x_2)$ is a hash, and the feature and class labels act as a hash key. When $h(x_1, x_2)$ is sparse, it is sometimes called a bag.

"Bag of Features" methods are increasingly used for learning and recognition.

Histograms and the Curse of Dimensionality

Computers and the Internet make it possible to directly apply histograms to very large amounts of data, and to consider very large feature sets. For such applications it is necessary to master the size of the histogram and the quantity of data.

Assume a feature vector $\vec{X}$, composed of D features, where each feature has one of N possible values.

The histogram "capacity" is the number of cells $Q=N^D$. Obviously, this grows exponentially with D. It is often convenient to reason in powers of 2 here.

Note 2^{10}=Kilo, 2^{20}=Meg, 2^{30}=Giga, 2^{40}=Tera, 2^{50}=Peta,

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}

In this case, the RMS error between a histogram and the underlying density is

$$E_{RMS}\,(h(X)-P(X)) = O(Q/M).$$

As a rule, it is recommended to have 10 samples per cell. $M \geq 10\,Q$.
The worst case occurs when the true underlying density is uniform.

For example, for D=5 features each with N = 32 values, the histogram has 1 Meg cells and you need 10 Meg of data.

For D= 6 features with N=64 values, h() has 1 Gig of cells and you need 10 Giga of samples.

For higher numbers of values or features, it is more convenient to work with probability densities.