Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2012/2013

Lesson 13 3 April 2013

Bayesian Recognition and Reasoning

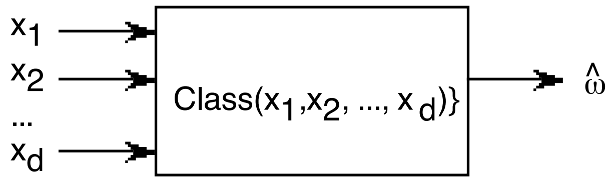
Sources Bibliographiques :
"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

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 $E \in C_k$
M_k	Number of examples for the class k. (think M = Mass)
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{\vec{x}_m^k\}$	A set of M_k examples for the class k. $\{\vec{x}_m\} = \bigcup_{k=1,K} \{\vec{x}_m^k\}$

Bayesian Classification

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 E belongs to class k:

$$\omega_k = E \in C_k$$

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

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

A fundamental tool for this is Baye's rule.

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

Supervised Learning

We will use a set of labeled "training set" of samples to estimate the probabilities $p(\vec{X})$, $p(\vec{X} \mid \omega_k)$, and $P(\omega_k)$. This is referred to as "supervised learning".

Assume that we have K classes.
For each class we have a set of M_k sample events $S_k = \{\vec{x}_m^k\}$.

The union of the training samples for each class gives us our training set:

$$S = \{\vec{x}_m\} = \bigcup_{k=1,K} \{\vec{x}_m^k\} \text{ composed of } M = \sum_{k=1}^{K} M_k \text{ samples (think M = Mass)}$$

In the simplest cases, we can use histogram (tables of frequencies) to represent the probabilities.

Illustrating Baye's Rule with Histograms

For simplicity, consider the case where D=1 with x is a natural number, $x \in [1, N]$, The same techniques can be made to work for real values and for symbolic values.

We need to represent $p(\vec{X})$, $p(\vec{X} \mid \omega_k)$, and $P(\omega_k)$.

Assume a training set $\{x_m\}$ of features from M events, such that $x \in [1, N]$composed of K subsets $\{\vec{x}_m^k\}$ of examples for each class k, with M_k examples in each subset.

$$\{\vec{x}_m\} = \bigcup_{k=1,K} \{\vec{x}_m^k\} \text{ and of } M = \sum_{k=1}^{K} M_k$$

We can build a table of frequency for the values of X. We allocate a table of N cells, and use the table to count the number of times each value occurs:

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

Then the probability that a random sample $X \in \{x_m\}$ from this set has the value x is then

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

Similarly if we have K classes, each with a set of M_k training samples $\{x_m^k\}$.
then we can build K histograms, each with N cells.

$$\forall k: \forall m=1, M: \ h_k(x_m) := h_k(x_m) + 1$$

Then

$$p(X = x \mid \omega_k) = \frac{1}{M_k} h_k(x)$$

The combined probability for all classes is just the sum of the histograms.

$$h(x) = \sum_{k=1}^{K} h_k(x) \text{ and then as before, } p(X = x) = \frac{1}{M} h(x)$$

$P(\omega_k)$ can be estimated from the relative size of the training set.

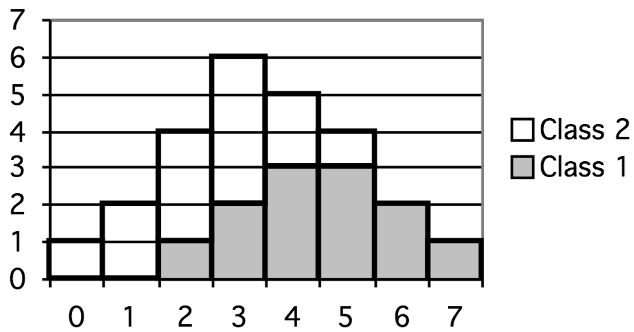
$$p(E \in C_k) = p(\omega_k) = \frac{M_k}{M}$$

Baye's Rule as a Ratio of Histograms

Note that this shows that the probability of a class is just the ratio of histograms:

Thus $$p(\omega_k \mid x) = \frac{p(x \mid \omega_k) p(\omega_k)}{p(x)} = \frac{\frac{1}{M_k} h_k(x) \frac{M_k}{M}}{\frac{1}{M} h(x)} = \frac{h_k(x)}{h(x)}$$

for example, when K=2



For example, observe that $p(\omega_1 \mid x=2) = 1/4$

Reminder. Using Histograms requires two assumptions:

1) that the training set is large enough ($M > 8\ Q$, where $Q=N^D$), and
2) That the observing conditions do not change with time (stationary),

We also assumed that the feature values were natural numbers in the range [1, N]. this can be easily obtained from any features.

When X is a vector of D features.

When X is a vector of D features each of the components must be normalized to a bounded integer between 1 and N. This can be done by individually bounding each component, x_d.

Assume a feature vector of D values $\vec{x}$

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_D \end{pmatrix}$$

Given that each feature $x_d \in [1, N]$, allocate a D dimensional table
$h(x_1, x_2, \dots, x_D) = h(\vec{x})$.

The number of cells in $h(\vec{X})$ is $Q=N^D$.
As before,

$$\forall m=1, M \; : \; h(\vec{X}_m) = h(\vec{X}_m) + 1$$

Then:

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

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

Where Q is the number fo cells in h(X)
N is the number of values for each feature.
D is the number of features.

Example: Grades in Two Courses

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 $M \geq 8Q = 200$.

An example, consider the table as follows:

	$h(x_1,x_2)$	A	B	C	D	F	$r(x_2)$
		x_1					
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 $h(x_1,x_2)$

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

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*:*

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

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

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.

Histograms for non-Integer Features

Unbounded and real-valued features

If X is real-valued of unbounded, we must bound it to a finite interval and quantize it. We can quantize with a function such as "trunc()" or "round()". The function trunc() removes the fractional part of a number. Round() adds ½ then removes the factional part.

To quantize a real X to N discrete values : [1, N]
x_{min}
/* first bound x to a finite range */

If (x < x_{min}) then x := x_{min};
If (x > x_{max}) then x := x_{max};

$$n = round\left((N-1)\cdot\frac{x - x_{\min}}{x_{\max} - x_{\min}}\right) + 1$$

Symbolic Features

If the features are symbolic, h(x) is addressed using a hash table, and the feature and feature values act as a hash key. As before h(x) counts the number of examples of each symbol. When symbolic x has N possible symbols then

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

"Bag of Features" methods are increasingly used for learning and recognition.
The only difference is that there is no "order" relation between the feature values.

Bayesian Reasoning as Evidence Accumulation

Bayesian Reasoning is a widely used technique to validate or invalidate hypothesis using uncertain or unreliable information. With this approach, a hypothesis statement, H, is formulated and assigned a probability, P(H). As new evidence, E, for or against the hypothesis is obtained it is also assigned a probability P(E) as well as a probability that it confirms the hypothesis, P(E|H). Baye's rule is then used to update the probability of the hypothesis:

$$P(H \mid E) \leftarrow \frac{P(E \mid H)P(H)}{P(E)}$$

In Bayesian reasoning, this rule is applied recursively as new evidence is obtained.

Let us define E_i as a body of prervious evidence composed of i elements, and E as a new element of evidence. Then Bayes rule tells us that :

$$P(H \mid E, E_i) \leftarrow \frac{P(E \mid H, E_i)}{P(E)} P(H, E_i)$$

to which we can then add $E_{i+1} \leftarrow E \cup E_i$

In this formula, the prior probability P(H) is simply the previous estimate of the probability of the hypothesis given the previous evidence. P(H, E_i). However, because the evidence E is independent of previous evidence, E_i you will often see $P(E \mid H)$ in place of $P(E \mid H, E_i)$. This gives:

$$P(H \mid E) \leftarrow \frac{P(E \mid H)P(H)}{P(E)}$$