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.

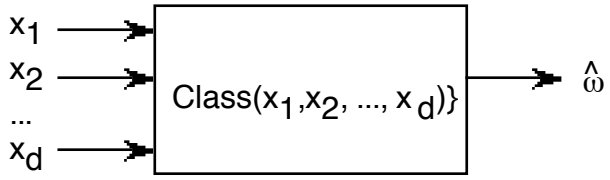
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

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

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

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

A fundamental tool for this is Baye's rule.

Baye's Rule :

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

We can use histograms to estimate the probability p(X=x).

Bayesian Probability

Baye's rule provides a method to accumulate evidence to reduce uncertainty.

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.

Using Histograms to Estimate Probability

When x is a natural number, $x \in [1, N]$, the application is obvious. However, the same technique works for real as well as symbolic values of x.

Given a training set $\{X_m\}$ of features from M events, such that $x \in [1, N]$, 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 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 ($M > 8\ Q$, 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.

Histogram Representation for a Bounded Integer

To use a histogram to build a non-parametric representation for numerical features the set of possible values for the feature must be finite. That is, each feature value must be represented by an integer x from a finite range:

$$x \in [x_{min}, x_{max}].$$

In many problems this occurs naturally. For example: the age, height, weight of a person, grades in a class, amount of change in a purse. In other cases, we can map the feature into a finite range.

For convenience, we will map features to integer values in the range $x \in [1, N]$,

If X is integer, with $x \in [x_{min}, x_{max}]$ we need only subtract x_{min}.

$$x := x - x_{min}.$$

Histograms for unbounded integer x.

If x is unbounded we must first bound it. We define bounds: x_{min} and x_{max}.
Then

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

Histograms for real x.

If X is real, we must quantize it 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 X to N discrete values :

For X real:

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

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

if $n > N$ then $n = N$.

This last line handles the rare case where $X = X_{max}$ and thus $n = N+1$.

Symbolic Features

If the features are symbolic, $h(x_1, x_2)$ is a hash, and the feature and class labels act as a hash key. If there are no order relations between the symbols, then $h(x_1, x_2)$ is called a bag.

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

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:

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

as we saw in the previous lecture, the average error depends on the ratio
$Q=N^D$ and M. : $E_{ms} \sim O(\frac{Q}{M})$

Illustrating Baye's 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 $Q = 5 \times 5 = 25$ 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:

		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.

Baye's Rule as a Ratio of Histograms

Suppose that we have 2 classes, k=1 and k=2, and that we observe a training set of M_1 events from class k=1: $\{\vec{X}_m^1\}$ and M_2 events from class k=2 $\{\vec{X}_m^2\}$

We assume that the feature vectors have D dimensions, each quantized to integer values in the range [1, N]. We assume stationary observing conditions with $M_1 \geq 8N^D$ and $M_2 \geq 8\ N^D$.

We build the histograms $h_1(\vec{x})$ and $h_2(\vec{x})$:

for m=1 to M_1 : $h_1(\vec{X}_m^1) := h_1(\vec{X}_m^1) + 1$

for m=1 to M_2 : $h_2(\vec{X}_m^2) := h_2(\vec{X}_m^2) + 1$

We also define $h(\vec{x}) = h_1(\vec{x}) + h_2(\vec{x})$ and $M = M_1 + M_2$

Thus, for a new observation, E, with features mapped to integers, then

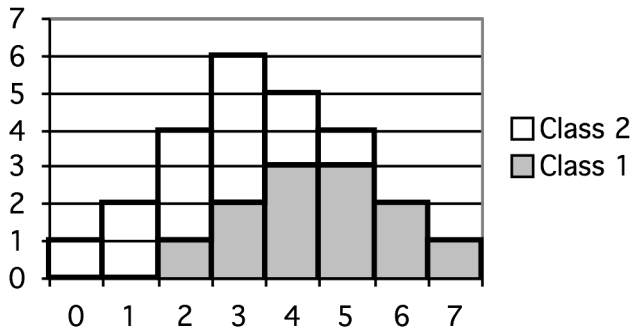
$$p(\vec{X}) = \frac{1}{M} h(\vec{x}) \quad \text{where } p(\vec{X}) \text{ is shorthand for } p(\vec{X} = \vec{x})$$

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

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

Thus $$p(\omega_1 \mid n) = \frac{p(\vec{X} \mid \omega_k) p(\omega_k)}{p(\vec{X})} = \frac{\frac{1}{M_k} h_k(\vec{x}) \frac{M_k}{M}}{\frac{1}{M} h(\vec{x})} = \frac{h_k(\vec{x})}{h(\vec{x})}$$

If D =1



For example, $p(\omega_1 \mid x=2) = ¼$

The probability of observing class k give feature x is $p(\omega_k|x) = h_k(x)/h(x)$

Variable size histogram cells

If the quantity of training data is too small, ie M < Q we can combine adjacent cells so as to amass enough data for a reasonable estimate.

Let us define the volume of each cell as 1.
Then the volume of the entire space is $Q=N^D$.

Suppose we merge V adjacent cells such that we obtain a combined sum of P. The volume of the combined cells would be V

$$P = \sum_{\vec{X} \in V} h(\vec{X})$$

The probability $p(\vec{X})$ for $\vec{X} \in V$ is $p(\vec{X}) = \frac{P}{MV}$

Suppose our samples $\{\vec{X}_m\}$ are drawn from a density $p(\vec{X})$.
If take a volume, V, from this density then

$$p(\vec{X}_m \in V) = \frac{P}{MV}$$

We can use this equation to develop two alternative non-parametric methods.

Fix V and determine P => Kernel density estimator.
Fix P and determine V => K nearest neightbors.

(note in most developments the symbol “K” is used for the sum the cells. This conflicts with the use of K for the number of classes. Thus we substitute the symbol P for the sum of adjacent cells).

Conclusions about Histograms

as a representation of probability, histograms have advantages and disadvantages.

Advantages
1) They have a fixed size, Q, independent of the quantity of data. It is not necessary to store the data.
2) They can be composed and used incrementally.

The disadvantage is that

1) Each feature must be quantized over a limited range of N values.
2) We need M >> Q data samples.
3) There are discontinuities at the boundaries of each cell.

Because the $M = \sum_{\vec{X}} h(\vec{X})$ we are sure that $\sum_{\vec{X}} p(\vec{X}) = 1$