Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2013/2014

Lesson 12 28 March 2014

Bayesian Reasoning and Recognition

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

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

Pattern Recognition

Recognize = Re + Cognize. To know again

Recognition is a fundamental ability for intelligence, and indeed for all life.
To survive, any creature must be able to recognize food, enemies and friends.

Recognition = assigning an observation to a class. This is also called "Classification".

Categorize is sometimes used in place of classify, generally when the categories have been determined automatically by machine learning.

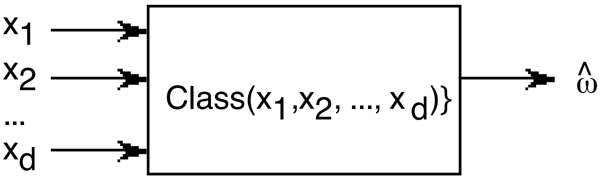
An observation (or event) E is provided by a sensor.
Generally and observation is described by a vector of properties called features, $\vec{X}$

For example: a medical scale that measures height and weight of a person. Each observation E is a vector $\vec{X}$ = (w,h). We can use this feature vector to guess the identity of the person. Let us refer to measures of the person k as the statement that the observation E belongs to the class C_k of observations of person k.

Features: observable properties that permit assignment of observations to classes.
A set of D features, x_d, are assembled into a feature vector $\vec{X}$

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

A classifier is a process that maps the observed properties $\vec{X}$ of an observation, E, to a class label, C_k. The result is a proposition: $\hat{\omega}_k =$ (E ∈ Class C_k)



Bayesian Reasoning and Classification

"Bayesian" refers to the 18th century mathematician and theologian Thomas Bayes (1702–1761), who provided the first mathematical treatment of a non-trivial problem of Bayesian inference. Bayesian inference was made popular by Simon Laplace in the early 19th century.

Bayesian inference can be used for reasoning and for recognition.
The rules of Bayesian inference can be interpreted as an extension of logic. Many modern machine learning methods are based on Bayesian principles.

With a Bayesian approach, the tests are designed to minimize the number of errors.

Let ω_k be the proposition that the event belongs to class k: $\omega_k = E \in T_k$

ω_k	Proposition that event E $\in$ the class k
$p(\omega_k) = p(E \in C_k)$	Probability that E is a member of class k

Given an observation E with properties $\vec{X}$, the decision criteria is

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

where $\omega_k \equiv E \in C_k$

To do this we need to define "probability" and "conditional probability" (given).

Once we are clear on the definition of probability, the meaning of conditional probability will be provided by Bayes Rule:

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

Probability

There are two possible definitions of probability that we can use for reasoning and recognition: Frequentialist and Axiomatic.

Probability as Frequency of Occurrence

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

Suppose we have M observations of random events (or observations) $\{E_m\}$, 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:

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

If we make new observations under the same 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 the size of the set of observations, M, grows larger. This is called the sampling error.

The sampling error is formally defined as the difference between calculated statistic and a parametric model. The parametric model is assumed to be true. Most often a Normal Density model is used.

The sampling error is generally inversely proportionally to M, the total number of observations

$$E_s \sim O(\frac{1}{M})$$

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 (or observation), let S be the set of all events, and let C_k be set of events that belong to class k with K total classes. $S = \bigcup_{k=1,K} C_k$

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

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

This can be very useful if we have some way to estimate the relative "likelihood" of different propositions, say $L(\omega_k)$.

We can convert a likelihood to probability by normalizing so that the sum of all likelihoods is 1. To do this we simply divide by the sum of all likelihoods:

$$P(\omega_k) = \frac{L(\omega_k)}{\sum_{k=1}^{K} L(\omega_k)}$$

Thus with axiomatic probability, any estimation of likelihood for the statement ω_k can be converted to probability and used with Bayes rule.

Histogram Representation of Probability

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

A histogram is a table of "frequency of occurrence" h(). When we have K classes of events, we can build a table of frequency of occurrence for observations from each class $h(E \in C_k)$.
Similarly if we have M observations of a feature, x, and the feature can take on one of N possible values, {1, ..., N} we can construct a table of frequency of occurrence for the feature. h(x).

For symbolic values, such as class labels or symbolic features, 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 <- C_k.

We then count the frequency of example of the class.

$$\forall m=1, M \ : \text{if } E_m \in C_k \text{ then } h(k) := h(k) + 1;$$

After M events, given a new event, E, $P(E \in C_k) = P(k) = \frac{1}{M} h(k)$

Similarly, for feature X with N possible values, we count the fequency of each value X_m within M observations.

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

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

<u>Problem</u>: Given a feature X, with N possible values, how many observations, M, do we need for a reliable estimate of probability?
<u>Answer</u>: If the feature X has N possible values, then 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 $O(\frac{Q}{M})$

The RMS (root-mean-square) sampling error between a histogram and the underlying parametric density model is $E_{RMS}(h(X)-P(X)) = O(Q/M)$.
The worst case occurs when the true underlying density is uniform.

For most applications, $M \geq 10 Q$ (10 samples per "cell") is reasonable (less than 10% RMS error).

when reasoning in powers of 2 one can use : $M \geq 8 Q$ (less than 12% RMS error).

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 $P(\omega_k)$. We will estimate these from a "training" set.

Assume that we have K classes.
For each class we have a set of M_k sample events, described by a feature $x \in [1, N]$

$$S_k = \left\{\vec{x}_m^k\right\}.$$

Overall we have M events, representing observations of K classes, with M_k examples in each class.

Traiing Set: $\{\vec{x}_m\} = \bigcup_{k=1,K} \{\vec{x}_m^k\}$ and $M = \sum_{k=1}^{K} M_k$

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 for each of the K classes, each with a set of M_k training samples $\left\{x_m^k\right\}$.
then we can build a 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.

$h(x) = \sum_{k=1}^{K} h_k(x)$ 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.

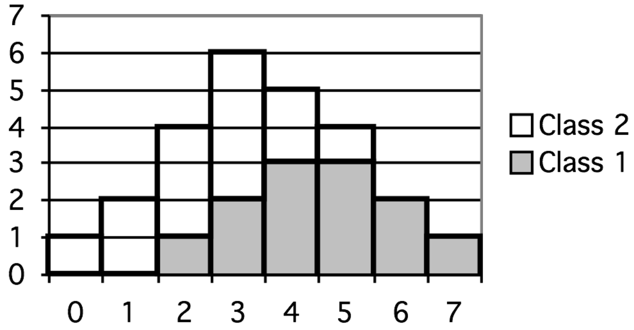
$$p(E \in C_k) = p(\omega_k) = \frac{M_k}{M}$$

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 $p(\omega_k \mid x) = \frac{p(x \mid \omega_k) p(\omega_k)}{p(x)} = \frac{\frac{1}{M_k} h_k(x) \frac{M_k}{M}}{\frac{1}{M} h(x)} = \frac{h_k(x)}{h(x)} = \frac{h_k(x)}{\sum_{k=1}^{K} h_k(x)}$

for example, when K=2



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

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.

As before, for K classes, where $h_k(\vec{x})$ is a histogram of feature vectors for class k:

$$p(\omega_k \mid \vec{x}) = \frac{p(x \mid \omega_k)p(\omega_k)}{p(x)} = \frac{\frac{1}{M_k} h_k(\vec{x}) \frac{M_k}{M}}{\frac{1}{M} h(\vec{x})} = \frac{h_k(\vec{x})}{h(\vec{x})} = \frac{h_k(\vec{x})}{\sum_{k=1}^{K} h_k(\vec{x})}$$

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.