Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2010/2011

Lesson 12 21 March 2012

Introduction to Bayesian 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 $E \in C_k$
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

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.

Two forms of recognition: Identify and Classify

Identify: To recognize an object or entity as an individual
Classify: To recognize an object or entity as a member of a class.

Categorize is sometimes used in place of classify.

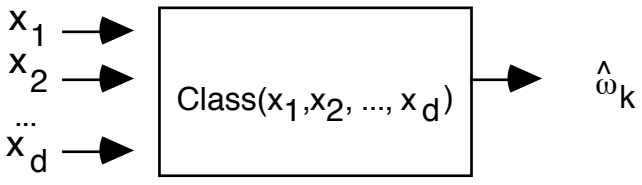
In this course we are interested in classifying observed events.
Classification is a process of associating an event to a class.
Each event is described by a set of features,

The event E is described by a vector of features, $\vec{X}$
Features are provided by an observation using sensors.
The observation returns a set of properties of the even.
These are called "Feature".

Features: observable properties that permit assignment of events 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 an event, E, to a class label, C_k, based on the features of the event. The result is the proposition $\omega_k = E \in \text{Class } C_k$



Bayesian 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 probability was made popular by Simon Laplace in the early 19th century.

The rules of Bayesian logic can be justified by requirements of rationality and consistency and interpreted as an extension of logic. Many modern machine learning methods are based on objectivist Bayesian principles.

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

For 2 class problems, false positives and false negatives count equally as errors, but can have different costs associated. This approach makes it possible to include the cost of error, which may not be the same for a false positive and a false negative.

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 $\vec{X}$, the decision criteria is

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

where $\omega_k \equiv E \in C_k$

The meaning of "given" is provided by Bayes Rule:

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

Applying Bayes rule for classification will require us to define probability.

Probability

Probability and Uncertainty

The core problem of recognition is uncertainty. One could even say that recognition is a problem of assigning signals to categories in the presence of uncertainty.

We can distinguish two separate kinds of uncertainties: Confidence and Accuracy (Precision).

Confidence: Freedom from doubt, belief in the truth of a proposition.
Accuracy: Reproducibility of a measurement.

Confidence concerns the truth of a statement. The proposition is generally formalized as a predicate (truth function). Predicates are generally defined a boolean truth functions (True or false). It is possible to define probabilistic truth functions.

Accuracy concerns a selecting an entity from an ordered set. Generally there is some order between the possible values with an associated distance metric. The accuracy refers to the size of a subset of possible values or the distance spanned by possible values.

In popular language, accuracy is often confused with precision.
In informatics:

Accuracy is the degree to which a measurement can be reproduced.
Precision is the detail with which a measurement is represented.

For example, a measurement may be represented with 32 bits of precision, but be accurate to only 8 bits (1 part in 256).
In common usage, precision and accuracy are often used for the same concept.

Probability is a powerful tool for both Confidence and Accuracy.

Both confidence and precision may be addressed in using Bayesian probabilities.

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:

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

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

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

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 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.

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 $h(E \in C_k)$.

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 <- C_k

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

After M events, given a new event, E,

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

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).

The RMS (root-mean-square) error between a histogram and the underlying density 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.
when reasoning in powers of 2 one can use : $M \geq 8\,Q$

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}

For example, for D=4 features each with N = 32=2^5 values, the histogram has $2^{4x5} = 2^{20}$ = 1 Meg cells and you need 8 Meg = 2^{23} samples of data.

For D= 5 features with N=64=2^6 values, h() has $2^{5x6} = 2^{30}$ = 1 Gig of cells and you need 2^{33} = 8 Giga of samples.

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