Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2014/2015

Lesson 11 18 March 2015

Bayesian Reasoning and Recognition

Bayesian Reasoning and Recognition

Baye's rule provides a unifying framework for pattern recognition and for reasoning under uncertainty. An important property is that this approach provides a framework for machine learning.

"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.

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

Bayes rule is expressed in terms of relative probabilities. Thus we need to review the definition of probability to provide a solid foundation.

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.

Assume that we have some form of process that generates events belonging to one of K classes. The event for each class is said to be "random". This means that the exact class can not be predicted in advance.

Such a process is commonly used to model sensors that make observations of the world. We will often refer to events as "observations".

Suppose we have a set of N observations of random events $\{E_n\}$, for which N_k of these events belong to the class k. The probability that one of these observed events from the set $\{E_n\}$ belongs to the class k is is the relative frequency of occurrence of the class:

$$P(E \in C_k) = \frac{N_k}{N}$$

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, N, grows larger. This is called the sampling error. We will return to this later.

Probability = relative frequency of occurrence.

A frequency based definition is easy to understand and can be used to build practical systems. It can also be used to illustrate basic principles. However it is possible to generalize the notion of probability with an axiomatic definition. This will make it possible to define a number of analytic tools.

Axiomatic Definition of probability

An axiomatic definition of probability makes it possible to apply analytical techniques to the design of reasoning an recognition 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 the subset of events that belong to class k with K total classes.

$S = \bigcup_{k=1,K} C_k$ is the set of all events.

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 $C_i \cap C_j = \varnothing$: $p(E \in C_i \cup C_j) = p(E \in C_i) + p(E \in C_j)$

Any function, *P(-)* that obeys these 3 postulates can be used as a probability.

An axiomatic definition of probability can be very useful if we have some way to estimate the relative "likelihood" of different propositions.

Let us define ω_k as the proposition that an event E belongs to class C_k : $\omega_k \equiv E \in C_k$

The likelihood of the proposition, $L(\omega_k)$, is a numerical function that estimates of its relative "plausibility" or believability of the proposition. Likelihoods do not have to obey the probability postulates.

We can convert a likelihood into a 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. This is fundamental for Bayesian reasoning and for Bayesian recognition.

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(-)*.

Suppose 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 N observations of an event, and the event can be from K possible classes, k=1,…, K. Suppose that the classes

We can construct a table of frequency of occurrence for the class. h(k).

$\forall n=1, N$: if $E_n \in C_k$ then h(k) := h(k) + 1;

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

Probabilities of Numerical Properties.

The notion of probability and frequency of occurrence are easily generalized to describe the likelihood of observed properties of objects. Such properties are often called "features" in the literature on pattern recognition.

For example, consider the height, measured in cm, of people present in this lecture today. Let us refer to the height of each student n, as a "random variable" X_n. X is "random" because it is not known until we measure it.

We can generate a histogram, h(x), for the N students present :
We first allocate a table h(), of, say 300 cells, for heights between 1 and 300 cm.
We then count the number of times each height occurs

$\forall n=1, N$: $h(X_n) := h(X_n) + 1$;

After counting the heights we can make statements about the population of students. For example, the relative likelihood of height that a random student has a height of X=180cm is

$$L(X=180) = h(180)$$

This is converted to a probability by normalizing so that the values of all likelihoods sum to 1 (axiom 2).

$$P(X = x) = \frac{1}{N} h(x)$$

We can use this to make statements about the population of students in the class (sample set) such as

1) The average height of a member of the class is:

$$\mu_x = E\{X_n\} = \frac{1}{N} \sum_{k=1}^{K} h(k) \cdot k = \frac{1}{N} \sum_{N=1}^{N} X_n$$

Where k is the set of possible heights in cm. (note that average is the first moment, or center of gravity of the histogram).

2) The largest height in the class is

$$x_{\max} = \arg\max_{k} \{h(k)\}$$

For symbolic features, such as gender or country of origin, the table *h()* can be implemented as a hash table, using the symbols as a key.

Alternatively, we can map the features onto K natural numbers k <- X_n.

However symbolic features do not necessarily have an order (as with numbers). Thus we cannot define an "average" value for a symbolic feature.
We CAN define a most likely value.

Probabilities of Vector properties.

We can also generalize to multiple properties. For example, each person in this class has a height, weight and age. We can represent these as three integers x_1, x_2 and x_3. Thus each person is represented by the "feature" vector $\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$.

We can build up a 3-D histogram, $h(x_1, x_2, x_3)$, for the N persons in this lecture as:

$$\forall n = 1, N : h(\vec{X}_n) = h(\vec{X}_n) + 1$$

or equivalently: $\forall n=1, N : h(x_1, x_2, x_3) := h(x_1, x_2, x_3) + 1;$

and the probability of a specific vector is $P(\vec{X} = \vec{x}) = \frac{1}{N} h(x_1, x_2, x_3)$

Number of samples required

Problem: Given a feature *X*, with K possible values, how many observations, N, do we need for a reliable estimate of probability?

Answer:
If the feature X has V possible values, then h(x) has Q = V cells.
If $\vec{X}$ is a vector of D features, each with K values then $Q = V^D$
Q is the histogram "capacity".

For N observations, in the worst case the sampling error is proportional to $O(\frac{Q}{N})$

This can be estimated by comparing the observed histograms to an ideal parametric model of the probability density or by comparing histograms of subsets samples to histograms from a very large sample.

The RMS (root-mean-square) sampling error between a histogram and the underlying parametric density model is $E_{RMS} = [E\{(h(X)-P(X))^2\}]^{1/2} = O(\frac{Q}{N})$.

The worst case occurs when the true underlying density is uniform.
For most applications, $N \geq 8\,Q$ (8 samples per "cell") is reasonable
(less than 12% RMS error).

Bayes Rule and Conditional probability.

Bayes Rule gives us a tool to determine "conditional probability".

Conditional probability is the probability of an event given that another event has occurred.

Conditional probability is one of the most fundamental and one of the most important concepts in probability theory.

Conditional probabilities require careful interpretation.
For example, there need not be a causal or temporal relationship between the events.
Conditional probability measures "correlation" or "association" not causality.

Consider two classes of events A and B.
Consider the probability that an event E belongs to A: $P(E \in A)$.
For convenience we will note this as P(A). (also sometimes written $P(\omega_a)$)

Similarly, consider the probability that an event belongs to the union $A \cup B$.
We will note this as $P(A \cup B)$

Conditional probability can be defined as

$$P(A \mid B) = \frac{p(A \cap B)}{P(B)}$$

Equivalently, conditional probability can be defined as

$$P(A \mid B)P(B) = P(A \cap B)$$

However, because set union is commutative:

$$P(A \mid B)P(B) = P(A \cap B) = P(B \cap A) = P(B \mid A)P(A)$$

The relation

$$P(A \mid B)P(B) = P(B \mid A)P(A)$$

is Bayes rule.

We can apply this to classes of events, as well as to the properties of objects (symbolic or numeric).

Consider the case where we have K classes of objects, C_k, where each object is characterized by a numerical feature X.

Suppose we have a set of N_k sample observations for each class, $S_k = \{X_n^k\}$.

We can use this set as a "training set" of samples to construct a histogram, *h(k, x)*.

h(k, x) is the joint probability of observing an event from class C_k it value X.

$$P((E \in C_k) \wedge (X=x)) = h(k, x)$$

This can be used to estimate the probabilities of observing a class member of class C_k given the observation of a feature vector $\vec{X}$.

$$P(E \in C_k \mid X=x) P(X=x) = h(k, x) = P(X=x \mid E \in C_k) P(E \in C_k)$$

$E \in C_k$ is commonly written as ω_k. It is common to simplify X=x as simply X.
This gives

$$P(\omega_k \mid x) P(x) = h(k, x) = P(x \mid \omega_k) P(\omega_k)$$

and thus $P(\omega_k \mid X) = \dfrac{P(\omega_k \mid X)P(\omega_k)}{P(X)}$

This also works for vector features and gives:

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

and $P(\omega_k \mid \vec{X}) = \dfrac{P(\omega_k \mid \vec{X})P(\omega_k)}{P(\vec{X})}$

Baye's Rule as a Ratio of Histograms

These probabilities can be estimated from the training samples, for example using histograms.

Consider an example of K classes of objects where objects are described by a an object property (or feature), *X*, with V possible values.
Assume that for each of the K classes, we have a "training set" of N_k samples $\{x_n^k\}$.

For each class k, we allocate a histogram, $h_k()$, with *V* cells and count the values in the training set.

$$\forall k: \forall n=1, N_k: \; h_k(X_n) := h_k(X_n) + 1$$

Then the probability of of observing a value X in the training set is

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

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

$$h(X) = \sum_{k=1}^{K} h_k(X) \text{ and } N = \sum_{k=1}^{K} N_k(X)$$

Thus

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

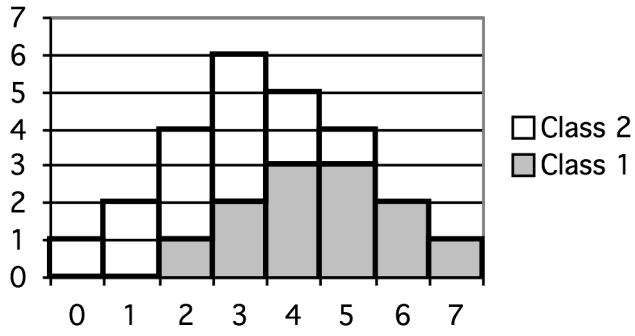
$P(\omega_k)$ can be estimated from the relative size of the training set.

$$p(E \in C_k) = p(\omega_k) = \frac{N_k}{N}$$

the probability that an observation E, with feature X belongs to class K is

$$p(\omega_k \mid X) = \frac{p(X \mid \omega_k) p(\omega_k)}{p(X)} = \frac{\frac{1}{N_k} h_k(X) \frac{N_k}{N}}{\frac{1}{N} h(X)} = \frac{h_k(X)}{h(X)} = \frac{h_k(X)}{\sum_{k=1}^{K} h_k(X)}$$

To illustrate, consider an example with 2 classes (K=2) and where X can take on 8 values (V=8, D=1).



If we observe an unknown object E with property X=2, then $p(\omega_1 | X=2) = ¼$

Note: Using Histograms requires two assumptions.
1) that the training set is large enough (N > 8 Q, where $Q=V^D$), 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, V]. this can be easily obtained from any features.

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.

Unbounded and real-valued features

If X is real-valued and unbounded, we can bind 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]

/* first bound x to a finite range */

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

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

Alternatively, we can present $p(\vec{X})$, $p(\vec{X} \mid \omega_k)$ as Probability Density Functions.

Example: Grades for students from 3 countries

Suppose we have a class with 38 students from 3 countries, attending the same class.
Let us call the countries C_k for k=1, 2, 3
Assume 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, $h(x_1, x_2)$, with Q= 5 x 3 = 15 cells.

Each student from each country is an event with feature $(x_1, x_2) = (C_k, G)$

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

Question: How many students are needed to fill this table?
Answer $N \geq 8Q = 120$.

An example, consider the table as follows:

		x_1			
	$h(x_1,x_2)$	C_1	C_2	C_3	$r(x_2)$
x_2	A	1	1		2
	B	4	3	1	8
	C	6	3	2	11
	D	4	1	1	6
	F	1			1
	$c(x_1)$	16	8	4	38

Any cell, (x_1, x_2) represents the probability that a student from X_1 got grade X_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 $h(x_1,x_2)$

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

for example $c(x_1=C_2) = 8,\ r(x_2=A) = 2,\ h(C_2,A) = 1$

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: $p(x_1 = C_2) = \sum_{x_2=A,B,\ldots,F} p(x_1 = C_2, x_2) = \frac{1}{N} \sum_{x_2=A,B,\ldots,F} h(C_2,x_2) = \frac{c(C_2)}{N} = \frac{8}{38}$

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 as ratio of histograms

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.

$$p(X_1 = C_1 \mid X_2 = B) = \frac{h(C_1,B)}{\sum_{x_1} h(x_1,B)} = \frac{4}{8} \text{ and } p(X_2 = D \mid X_1 = C_1) = \frac{h(D,C_1)}{\sum_{x_2} h(D,x_2)} = \frac{4}{6}$$

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

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

or more commonly written: $p(X_1 \mid X_2) = \frac{p(X_2 \mid X_1) \cdot p(X_1)}{p(X_2)}$