Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2015/2016

Lesson 7 2 March 2016

Bayes Rule and Conditional Probability

Classification

Our problem is to build a box that maps the observation, $\vec{X}$ into an estimate $\hat{C}_k$ of the class C_k from a set of K possible classes.

Where $\hat{\omega}_k$ is a statement that asserts that $\vec{X}$ is a member of class $\hat{C}_k$, one of the K possible classes. $\hat{C}_k \in \{C_k\}$

With a generative approach to classification, we will design our box to minimize the number of mistakes we will look for the most likely class, $\hat{C}_k \in \{C_k\}$

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

Our primary tool for this will be Baye's Rule : $P(\omega_k \mid \vec{X}) = \dfrac{P(\vec{X} \mid \omega_k) P(\omega_k)}{P(\vec{X})}$

To apply Baye's rule, we require a representation for the probabilities $P(\vec{X} \mid \omega_k)$, $P(\vec{X})$, and $P(\omega_k)$.

In today's lecture we will review the definitions for Bayes Rule and for Probability.

Bayes Rule and Conditional Probability

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.

Baye's Rule

Bayes Rule gives us a tool to reason with conditional probabilities.

Conditional probability is the probability of an event given that another event has occurred. Conditional probability measures "correlation" or "association".
Conditional probability does not measure causality.

Consider two independent classes of events A and B such that $A \cap B \neq \varnothing$
Let P(A) be the probability that an event $E \in A$
Let P(B) be the probability that an event $E \in B$ and
Let P(A, B) be the probability that the event is in both A and B.

We can note that $P(A, B) = P((E \in A) \wedge (E \in B)) = P(E \in A \cap B)$

Conditional probability can be defined as

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

Equivalently, conditional probability can be defined as

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

However, because set union is commutative:

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

The relation $P(A,B) = P(A \mid B)P(B) = P(B \mid A)P(A)$ is Baye's rule.

This can be generalized to more than 2 classes:

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

We can apply this to recognizing classes based on observed properties.
We will also use this to reason among alternative hypotheses.

But first we need to be clear about what we mean by probability.

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 sensor that generates observations belonging to one of K classes, C_k. The class for each observation is "random". This means that the exact class cannot be predicted in advance.

Suppose we have a set of M observations $\{E_m\}$, for which M_k of these events belong to the class C_k. The probability that one of these observed events from the set $\{E_m\}$ belongs to the class C_k. is the relative frequency of occurrence of the class in the set $\{E_m\}$.

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

If we make new observations under the same condition, 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.

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 and recognition systems. Only three postulates (or axioms) are necessary:

In the following, let E be an observation (even), 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.

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

Any function *P(-)* that obeys the following 3 axioms can be used as a probability:

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

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 set of likelihoods into probabilities 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" $h(-)$ for a class of events.
Histograms are typically constructed for numerical arguments $h(k)$ for $k=1, K$.

However many modern programming languages provide a structure called a "hash", allowing the construction of tables using symbolic indexes.

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 M observations of an event, $\{E_m\}$ 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).

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

Then for any event in $E \in \{E_m\}$: $P(E \in C_k) = P(k) = \frac{1}{M} h(k)$

Assuming the observation conditions are the same, given a new event, E,

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

Probabilities of Numerical Properties.

The notion of probability and frequency of occurrence are easily generalized to describe the likelihood of numerical properties (features), X, observed by sensors.
For example, consider the height, measured in cm, of people present in this lecture today. Let us refer to the height of each student m, as a "random variable" X_m. X is "random" because it is not known until we measure it.

We can generate a histogram, $h(x)$, for the M students present.
For convenience we will treat height as an integer from the range 1 to 300.
We will allocate a table $h(x)$, of 250 cells.

The number of cells is called the capacity of the histogram, Q.

We then count the number of times each height occurs in the class.

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

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).

$$P(X = x) = \frac{1}{M} h(x) \quad \text{where} \quad M = \sum_{x=1}^{250} h(x)$$

We can use this to make statements about the population of students in the class:

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

$$\mu_x = E\{X_m\} = \frac{1}{M} \sum_{m=1}^{M} X_m = \frac{1}{M} \sum_{x=1}^{250} h(x) \cdot x$$

Note that the average is the first moment, or center of gravity of the histogram.

2) The variance is the square of the average difference from the mean:

$$\sigma_x^2 = E\{(X_m - \mu_x)^2\} = \frac{1}{M} \sum_{m=1}^{M} (X_m - \mu_x)^2 = \frac{1}{M} \sum_{x=1}^{250} h(x) \cdot (x - \mu_x)^2$$

The average difference from the mean, σ_x, is called the "standard deviation", and is often abbreviated "std." In french we call this the "écart type".

Average and variance are properties of the sample population.

Histograms with integer and real valued features

If X is an integer value then we need only bound the range to use a histogram

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

Then allocate a histogram of $N=x_{max}$ cells.

We may, for convenience, shift the range of values to start at 1, so as to convert integer x to a natural number:

$n := x - x_{min}+1$

This will give a set of $N = x_{max} - x_{min} + 1$ possible values for X.

If X is real-valued and unbounded, we can limit it to a finite interval and then 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 natural numbers : [1, N]

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

Many modern languages contain a structure called a hash.
This is a table *h(x)* that uses a symbolic value, x, an address.

We can use a hash to generalize histograms for use with symbolic features.
The only difference is that there is no "order" relation between the feature values.

As before *h(x)* counts the number of examples of each symbol.
For example with M training samples drawn from N symbolic values.

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

Then for any random X,

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

Note that symbolic features do not have an order relation.
Thus we cannot define an "average" value for a symbolic feature.
We CAN estimate a most likely value.

$$\hat{x} = \arg\max_{x} \{h(x)\}$$

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 M persons in this lecture as:

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

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

and the probability of a specific vector is $P(\vec{X} = \vec{x}) = \frac{1}{M} h(\vec{x})$

When each of the D features can have N values, the total number of cells in the histogram will be $Q = N^D$

Number of samples required

Problem: Given a feature *x*, with N possible values, how many observations, M, do we need for a histogram, *h(x)*, to provide a reliable estimate of probability?

The worst case Root Mean Square error is proportional to $O(\frac{Q}{M})$.

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. Let *p(x)* be a probability density function. The RMS (root-mean-square) sampling error between a histogram and the density function is

$$E_{RMS} = \sqrt{E\left\{\left(h(x) - p(x)\right)^2\right\}} \approx O(\frac{Q}{M})$$

The worst case occurs for a uniform probability density function.

For most applications, $M \geq 8\,Q$ (8 samples per "cell") is reasonable (less than 12% RMS error).