Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 Winter Semester 2018
Lesson 3 14 February 2018

Bayes Rule and Probability

Notation

x	A variable
X	A random variable (unpredictable value). an observation.
M	The number of possible values for *X*
$\vec{x}$	A vector of D variables.
$\vec{X}$	A vector of D random variables.
D	The number of dimensions for the vector $\vec{x}$ or $\vec{X}$
C_k	The class k
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that $X \in C_k$
$P(X \in C_k)$	Probability that the observation X is a member of the class k.
M_k	Number of examples for the class k.
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{\vec{X}_m\}$	A set of training samples
$\{y_m\}$	A set of indicator vectors for the training samples in $\{\vec{X}_m\}$
$p(X)$	Probability density function for a continuous value X

Bayes Rule

Bayes rule provides a unifying framework for pattern recognition and for reasoning about hypotheses under uncertainty.

"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 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".

Consider two classes of events A and B.
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 $P(A, B) = P((E \in A) \wedge (E \in B)) = P(E \in A \cap B)$
Or even that: $P(A, B) = P(A \wedge B) = P(A \cap B)$
Conditional probability can be defined as

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

Equivalently, conditional probability can be defined as

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

Set union is commutative, giving $P(A,B) = P(B,A) = P(B \mid A)P(A)$

This gives the common definition of Bayes Rule:

$$\boxed{P(A \mid B)P(B) = P(B \mid A)P(A)}$$

This can be generalized to more than 2 classes:

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

To use these tools we need to be clear about what we mean by probability.

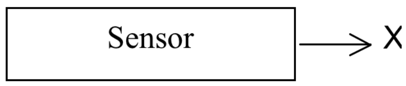
Probability

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 $\{X_m\}$, for which M_k of these events belong to the class C_k. The probability that one of these observed events from the set $\{X_m\}$ belongs to the class C_k is the relative frequency of occurrence of the class in the set $\{X_m\}$.

The probability that X_m belongs to C_k is $P(X_m \in C_k) = \frac{M_k}{M}$

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.

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 X be an observation (or event), let S be the set of all possible observations, and let C_k be the subset of observations that belong to class k from one of K classes, $\{C_k\}$

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

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

For any observation (event) E:
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 $C_i \cap C_j = \emptyset$: $P(E \in C_i \bigcup C_j) = P(E \in C_i) + P(E \in C_j)$

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 observation 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.

Assuming that $L(\omega_k)$ obeys axioms 1 and 3, we can convert 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:

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

We will use three different representations for probability:
Distribution Tables, Histograms and Density functions.

Probability Distribution Tables

A **Probability Distribution Table** that gives the relative frequency of occurrence for all possible values of a property (a feature) for a set of observations.

Suppose that we have a set of M observations that can be divided into N subsets, such that the subsets are (1) Mutually Exclusive and (2) Complete.

For example, a set of M people can be divided into subsets defined by eye color: C={Blue, Green, Brown}. This set is (1) Mutually Exclusive and (2) Complete.

Mutually Exclusive: A person can only belong to a single subset
Complete: A person must belong to one of the subsets.

These subsets represent observable properties of an event or observation. These are commonly called features. The values are called Feature Values.
Features can be Boolean, symbolic or numeric (integer or real)

A Probability Distribution Table that gives the relative frequency of occurrence for each value of a feature for a set of observations.

We will start by illustrating this with symbolic features.

Consider a set of people.
Let C represent the Eye Color C={blue, green, brown}, N_c=3.

Let *h(C)* be a counter for the value of C. *h(C)* is initially 0.
This can be easily implemented as a "map" that associates a key with a value.
The keys are the subset labels: {Blue, Green, Brown}
The values are the number of events with the value. *h(C)*
Capital C is the random variable for the set, lower case c is a specific value of C.

For each person E in the set S: if *E* is *c* then $h(c) \leftarrow h(c) + 1$
Formally:

$$\forall E \in S : E \in c \Rightarrow h(c) \leftarrow h(c) + 1;$$

Note that because each person can have one and only one feature value:

$$M = \sum_{c \in C} h(c)$$

Then probability distribution table gives the probability that a person E in the set S has the eye color c. This can be computed from:

$P(E \in c) = \frac{1}{M} h(c)$ This is commonly written: $P(C) = \frac{1}{M} h(C)$

Note that to be a valid probability, the values must sum to 1:

$$1 = \sum_{c \in C} P(c)$$

The most probable feature value is the feature value with the highest probability

$$\hat{c} \leftarrow \arg\text{-}\max_{c \in C} \{P(c)\}$$

This is a property of the set S and not the individuals, E, of the set.

Joint Probability Distributions Tables

Distribution tables can be generalized to multiple classes.

For example, the persons in the set S can have gender as well as Eye color.
Let G represent the Gender {Male, Female}. N_G=2
A joint distribution table counts the number of persons Eye Color, C, with a certain Gender, G

$$\forall E \in S \text{ IF } E \in c_1 \wedge E \in c_2 \text{ THEN } h(c_1, c_2,) \leftarrow h(c_1, c_2) + 1;$$

Then: $P(E \in C \wedge E \in G) = P(C,G) = \frac{1}{M} h(C, \text{G})$

The complete table must sum to 1. $\sum_{c \in C} \sum_{g \in G} P(c,g) = 1$

We can eliminate a class from the table by summing a column:

$$P(C) = \sum_{g \in G} P(C,g)$$

All this can be generalized to multiple features. For three features A, B, C

$$p(A,B,C) = \frac{1}{M} h(A,B,C)$$

and

$$P(A,B) = \sum_{x \in C} P(A,B,x)$$

Graphically, probability distribution tables are displayed as:

G\C	Brown	Blue	Green
Male	0.3	0.2	0.1
Female	0.3	0.2	0.1

When reasoning with Boolean features, some authors omit the columns for False.

Conditional Probability Tables (CPT)

Bayes Rule provides a definition of conditional probability tables.

For a probability distribution *P(A,B)* the Conditional probability can be defined as

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

With multiple features; $P(A,B \mid C) = \frac{P(A,B,C)}{\sum_{x \in C} P(A,B,x)} = \frac{P(A,B,c)}{P(A,B)}$

For example, consider the Boolean values F=Fire and S=Smoke

P(Fire, Smoke) = P(Smoke|Fire) P(Fire)

P(Smoke|Fire)	Smoke	¬Smoke
Fire	0.7	0.3
¬Fire	0	1

Each row sums to one. Columns are independent.

Histograms: 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.

An important difference is that with numerical values, the values obey an order relation: $1 < 2 < 3$, and a distance metric. $\forall x : dist(x, x+1) = 1$

This can be summarized using the triangle inequality:

$$\forall x_1, x_2, x_3 : dist(x_1, x_3) = dist(x_1, x_2) + dist(x_2, x_3)$$

With symbolic values such as Blue, Green and Brown there is no metric for distance.

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 $X := x_{max}$;

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

Histogram 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

<u>Problem</u>: 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, $M \geq 8Q$ (8 samples per "cell") is reasonable
(less than 12% RMS error).

Probability Density Functions

A probability density function (PDF) is:
A probability density function *p(X),* is a function of a continuous variable *X* such that

1) *X* is a continuous real valued random variable with values between $[-\infty, \infty]$
2) $\int_{-\infty}^{\infty} p(X) = 1$

Note that *p(X)* is NOT a number but a continuous function.

A probability density function defines the relatively likelihood for a specific value of *X*. Because *X* is continuous, the value of *p(X)* for a specific *X* is infinitely small. To obtain a probability we must integrate over some range of *X*.
To obtain a probability we must integrate over some range V of X.
In the case of D=1, the probability that X is within the interval [A, B] is

$$P(X \in [A,B]) = \int_{A}^{B} p(x)dx$$

This integral gives a number that can be used as a probability.

Note that we use upper case $P(X \in [A,B])$ to represent a probability value,
and lower case *p(X)* to represent a probability density function.

Bayes Rule with probability density functions

Classification using Bayes Rule can use probability density functions

$$P(\omega_k \mid X) = \frac{p(X \mid \omega_k)}{p(X)} P(\omega_k) = \frac{p(X \mid \omega_k)}{\sum_{k=1}^{K} p(X \mid \omega_k)} P(\omega_k)$$

Note that the ratio $\frac{p(X \mid \omega_k)}{p(X)}$ IS a number, provided that $p(X) = \sum_{k=1}^{K} p(X \mid \omega_k)$

Probability density functions are easily generalized to <u>vectors of random variables</u>.
Let $\vec{X} \in R^D$, be a vector random variables.
A probability density function, $p(\vec{X})$, is a function of a vector of continuous variables

1) $\vec{X}$ is a vector of D real valued random variables with values between $[-\infty, \infty]$
2) $\int_{-\infty}^{\infty} p(\vec{x})d\vec{x} = 1$

We concentrate on the Gaussian density function.