Intelligent Systems: Reasoning and Recognition

James L. Crowley

MoSIG M1 Winter Semester 2020
Lesson 3 11 February 2020

Bayes Rule with Probability Distributions and Densities

Notation

x	A variable
X	A random variable (unpredictable value). an observation.
N	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

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 C_k in the set $\{E_m\}$.

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

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 E 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, $\{C_k\}$

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

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.

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 $P(A, B) = P((E \in A) \wedge (E \in B)) = P(E \in A \cap B)$
We write this as : $P(A, B) = P(A \wedge B) = P(A \cap B)$

Conditional probability can be defined as $P(A \mid B) = \dfrac{P(A,B)}{P(B)}$

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:

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

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 techniques to represent and compute probability.

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 (or events).

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. These subsets represent "values", *X*, for the observation. (sometimes called Feature Values).

Values are mutually exclusive and the set of values is complete.
A single observation has a unique value, *X*. The value of an observation must be from the set of possible values.

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

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 X represent the Eye Color *X*={blue, green, brown}, $N_c=3$.

Let *h(x)* be a table of 3 counters for the values of X. *h(x)* is initially 0.

The Table *h(x)* can be easily implemented as a "map" that associates a key with a value. The keys are the labels: {Blue, Green, Brown}
The values are the number of observations with each lable. *h(x)*
Capital *X* is the random variable for the set of labels, lower case x is a specific value of *X*.
Suppose that we have a set M people, such that $\{X_m\}$ is the eye color of person m.

For each person we will increment the table $h(X_m) \leftarrow h(X_m) + 1$

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

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

$$M = \sum_x h(x)$$

Then probability distribution table gives the probability that a person E_m has eye-color X. This can be computed from:

$$P(X_m = x) = \frac{1}{M}h(x) \quad \text{This is commonly written: } P(X_m) = \frac{1}{M}h(X_m)$$

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

$$1 = \sum_x P(x)$$

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

$$\hat{X} \leftarrow \arg\!-\!\max_x\{P(x)\}$$

This is a property of the set S and not of the individuals 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 C_m represent the Eye-color of person M. $C_m \in$ {Blue, Green, Brown}, N_c=3
Let G_m represent the Gender of the person $G_m \in$ {Male, Female}, N_G=2

A joint distribution table counts the number of persons Eye Color, C, with a certain Gender, G

$$\forall m = 1,M: \ h(C_m, G_m) \leftarrow h(C_m, G_m) + 1$$

Then: $$P(C_m = c \wedge G_m = g) = \frac{1}{M}h(c,g)$$

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:

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

Graphically, probability distribution tables are displayed as:

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

Note that the table must sum to 1.

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

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

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 question: R?: Will it Rain today?
Rain can also be predicted by Cloud-Cover with values Clear, Scatttered, Overcast.
CLR = < 1/8 Clouds, 1/8 Clouds ≤ SCT < 6/8 Clouds , OVC IF ≥ 6/8 Clouds

P(Rain|Clouds)	Rain	¬Rain
CLR	0	1
SCT	0.2	0.8
OVC	0.6	0.4

We can then compute P(Rain) = P(Rain|Clouds) P(Clouds)

In a Bayes Net this is noted as:

If we do not observe clouds, we can use statistics to estimate P(Clouds).
Even if someone observe clouds, the judgment of cloud cover is subjective and may have errors.

P(Clouds|Observation)	CLR	SCT	OVC
CLR	0.9	0.1	0.0
SCT	0.2	0.7	0.1
OVC	0	0.2	0.8

Thus we can use Conditional Probabilities to account for observation errror.

P(Cloud | Observation = SCT) may be a distribution : (0.1, 0.7, 0.2).

Rain can be predicted from atmospheric pressure P with values Low, Medium, High
Low if $P < 1008$ HP, Med if $1008 \leq P < 1018$, High IF $P \geq 1018$

P(Rain|Pressure)	Rain	¬Rain
Low	0.8	0.2
Medium	0.4	0.6
High	0	1

P(R) = P(R|P) P(P)

These can be combined into a joint table P(R | P, C)

P(R) = P(R | P, C) P(P) P(C)

We will develop models for diagnostic reasoning using distribution tables later in the course.

It is also possible to use Bayes rule with numerical properties.
For natural and integer numbers, we can use Histograms.
For Real numbers, we can use density functions.

Histograms for 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 X_m.

We can generate a histogram, $h(x)$, for the M students present.
For convenience we will treat height as an integer from the range 151 to 250.
We will allocate a table $h(x)$, of 100 cells. The size of the table is Q=100

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).

$$P(X = x) = \frac{1}{M} h(x) \quad \text{where} \quad M = \sum_{x=151}^{250} h(x)$$

However, for this to be valid we need to have more data than the number of histogram cells: M >> Q. The general rule is 1

Bayes Rule with a Ratio of Histograms

Consider an example of K classes of objects where objects are described by a feature, *X*, with N possible integer values from [1, N]. Assume that we have a "training set" of M samples $\{x_m\}$ along with indicator variables $\{y_m\}$ where the indicator variable is the class, k, for each training sample.

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

```
$\forall_{m=1}^{M}$ : $h(X_m) \leftarrow h(X_m)+1$
        IF $y_m = k$ THEN
                $h_k(X_m) \leftarrow h_k(X_m)+1;$
                $M_k \leftarrow M_k+1$
```

Then

$$P(X=x \mid X \in C_k) = P(X \mid \omega_k) = \frac{1}{M_k} h_k(x)$$

The histogram for all possible values is the sum of histograms:

$$P(X=x) = \frac{1}{M}h(x) = \frac{1}{M}\sum_k h_k(x)$$

and $P(\omega_k)$ can be estimated from the relative number of events in each class.

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

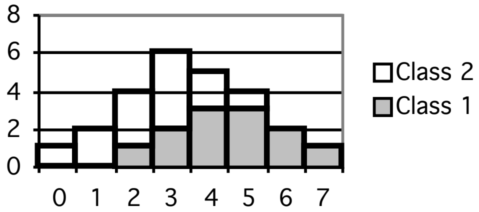
Bayes rule tells us: $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

giving: $$P(\omega_k \mid X) = \frac{P(X \mid \omega_k)P(\omega_k)}{P(X)} = \frac{\frac{1}{M_k}h_k(X)\frac{M_k}{M}}{\frac{1}{M}h(X)} = \frac{h_k(X)}{h(X)}$$

This can also be written as: $P(\omega_k \mid X) = \frac{h_k(X)}{\sum_{k=1}^{K} h_k(X)}$ because $h(X) = \sum_{k=1}^{K} h_k(X)$

The ratio of histograms can be represented by a lookup table. $P(\omega_k \mid X) = T(X)$

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



Note that having M >> Q is NECESSARY but NOT Sufficient.
Having M < Q is a guarantee of INSUFFICIENT TRAINING DATA.

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 ≥ 8 Q (8 samples per "cell") is reasonable
(less than 12% RMS error).

So what can you do if you do not have M >> Q ?
Adapt the size of the cell to the data!

Mean and Standard Deviation

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

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

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

Consider a table h(x) of the height of the students in this class.

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

$$\mu_x = E\{x_m\} = \frac{1}{M}\sum_{m=1}^{M} x_m = \frac{1}{M}\sum_{x=x_{\min}}^{x_{\max}} h(x)\cdot x$$

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

Histograms for Vectors of 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 $Q = N^D$

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

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

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

Probability density functions are easily generalized to vectors of random variables.
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.