Classification: Probabilistic Methods

We're going to begin our study of AI by looking at a task that is, on the one hand, easy to understand and yet, on the other hand, very widespread. The task is called *classification*.

1 Classification

In *classification problems*, the task is to decide to which of a predefined, finite set of categories an object belongs.

Note that the classes or categories are known in advance; there is a finite number of them; and, in fact, there is usually only a small number of them. Each class or category has a name, which we refer to as the *class label*.

We refer to the objects as *instances*.

A classification system (or classifier) receives a description of the instance to be classified; decides to which class the instance belongs; and outputs the appropriate class label. We will assume that instances are described by values for a set of attributes.

Here are some examples:

- In a game, one character has to decide whether another character is friend or foe, based on the appearance or behaviour of the other character.
- In a public building, a security system may allow access to private areas of the building to employees but not to visitors, determined by an analysis of images from a video camera.
- In an email system, a spam filter decides whether an incoming email is spam or ham, based on the email's content and formatting.
- In a language school, a decision agent must decide whether a newly-enrolled student will study at beginner, intermediate or advanced level, based on the student's performance in a language exam.

Classification may also be a subtask within other more complicated tasks. For example, to understand human language requires the ability to choose the correct meanings of ambiguous words; this is a classification task.

There are many, many ways to build classifiers. The first classification methods we are going to study use probabilities.

2 Probabilities

2.1 Probabilities and joint probabilities

In probability theory, we talk of *random variables*, $A, B, \ldots$. These are just variables that can take on different values. In our case, they will stand for features of our domain of interest (e.g. attributes of instances or the class of an instance). The values of the random variables can be of different types: Boolean (e.g. *sunny* $=$ true), numeric (e.g. *income* $= 60000$, *height* $= 1.73$) or symbolic (e.g. *pollen* $=$ *low*). For simplicity, we will restrict our attention to random variables whose values are discrete and finite.

Suppose A is a random variable whose possible values are $\{a_1, a_2, \ldots, a_k\}$. The *probability* that attribute A has value a_i is written $P(A = a_i)$. The probabilities must satisfy the following:

- $0 \leq P(A = a_i) \leq 1$

- $P(A = a_1) + P(A = a_2) + \ldots + P(A = a_k) = 1$

The *joint probability* that $A = a_i$ and $B = b_j$ is written $P(A = a_i, B = b_j)$ (and similarly for joint probabilities involving more than two variables). A variable can take on only one value so

- $P(A = a_i, A = a_j) = 0$ if $a_1 \neq a_j$

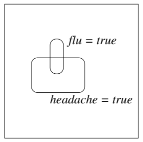
2.2 Conditional probabilities

The *conditional probability of* $A = a_i$ *given* $B = b_j$ is written $P(A = a_i | B = b_j)$. It is defined as follows:

$$P(A = a_i | B = b_j) = \frac{P(A = a_i, B = b_j)}{P(B = b_j)}$$

This definition can be used whenever $P(B = b_j) > 0$.

Suppose we have Boolean-valued variables *headache* and *flu*. And suppose $P(\textit{headache} = \text{true}) = 0.1$, $P(\textit{flu} = \text{true}) = 0.025$ and $P(\textit{headache} = \text{true}, \textit{flu} = \text{true}) = 0.0125$. Then, using the definition above, $P(\textit{headache} = \text{true} | \textit{flu} = \text{true}) = 0.5$. As a visualisation, think of each point in the following Venn diagram as a possible state of affairs. To reflect the idea that $P(\textit{headache} = \text{true}) = 0.1$, the region where *headache* $=$ true covers 0.1 of the whole area, and similarly for other probabilities/areas:



It could be said that $P(\textit{headache} = \text{true} | \textit{flu} = \text{true})$ measures the probability of *headache = true* relative to the reduced set of states in which *flu* $=$ true.

If we multiply through by $P(B = b_j)$, the definition of conditional probability can be rewritten as the following, known as the *product rule*:

$$P(A = a_i, B = b_j) = P(A = a_i | B = b_j) \times P(B = b_j)$$

2.3 Independence and conditional independence

$A = a_i$ is said to be *independent* of $B = b_j$ if the probability of $A = a_i$ is not influenced by whether $B = b_j$ holds. The definition is: $A = a_i$ and $B = b_j$ are *independent* if

$$P(A = a_i | B = b_j) = P(A = a_i)$$

or, equivalently,

$$P(B = b_j | A = a_i) = P(B = b_j)$$

or, equivalently,

$$P(A = a_i, B = b_j) = P(A = a_i) \times P(B = b_j)$$

The last of these will be useful to us.

Class Exercise. How might we show diagrammatically that *headache* $=$ true and *eyeColour* $=$ *blue* are independent?

Sometimes $A = a_i$ and $B = b_j$ will not be independent in any absolute sense but they may be independent when a third factor is fixed. I'm sure, for example, if we could obtain the probabilities, we would find that *shoeSize* $=$ *small* and *languageSkill* $=$ *low* are not independent: I'm sure we would find that if you have small feet then you are more likely to have low language skills! But this is a spurious relationship that depends on a third, though unmentioned, factor: age. Once age is fixed ('given'), then the dependence disappears. *shoeSize* $=$ *small* given *age* $=$ *young* and *languageSkill* $=$ *low* given *age* $=$ *young* are independent. This is called *conditional independence*.

The definition is: $A = a_i$ and $B = b_j$ are *conditionally independent* given $C = c_k$ if

$$P(A = a_i, B = b_j | C = c_k) = P(A = a_i | C = c_k) \times P(B = b_j | C = c_k)$$

This definition can be used whenever $P(C = c_k) > 0$.

3 Inference on a Joint Probability Distribution

3.1 What is a joint probability distribution?

Suppose we have two random variables $weather$ and $pollen$ whose sets of legal values are $\{cloudy, sunny\}$ and $\{low, medium, high\}$ respectively. Then we can show all the possible joint probabilities by a table with 6 rows. For example:

weather	*pollen*	Prob
cloudy	*low*	0.5
cloudy	*medium*	0.01
cloudy	*high*	0.01
sunny	*low*	0.2
sunny	*medium*	0.19
sunny	*high*	0.09

Each row in this table represents a different state of the world and so they are all mutually exclusive. The probabilities in the table must sum to 1.

3.2 Where do joint probability distributions come from?

1. Elicit them from human experts.
2. Compute them from simpler probabilities, e.g. using the product rule.
3. Learn them from data. If you have a large dataset, then you can estimate the probabilities:

$$P(\text{row}) = \frac{\text{number of records that match row}}{\text{total number of records}}$$

The following joint probability distribution, for example, was learned from a U.S. census database:

sex	*hours_worked*	*wealth*	Prob
female	40.5−	*poor*	0.25
female	40.5−	*rich*	0.03
female	40.5+	*poor*	0.04
female	40.5+	*rich*	0.01
male	40.5−	*poor*	0.33
male	40.5−	*rich*	0.1
male	40.5+	*poor*	0.13
meale	40.4+	*rich*	0.11

3.3 Inference using the joint probability distribution

You can obtain any other probability from the joint distribution. To find $P(E)$ for some expression E, you sum the probabilities in the rows that match E.

For example, $P(sex = male, wealth = poor) = 0.33 + 0.13 = 0.46$. Similarly, $P(wealth = poor) = 0.75$ (by summing the 4 rows in which *wealth* $=$ *poor*). And, having computed $P(sex = male, wealth = poor)$ and $P(wealth = poor)$, we can plug these probabilities into the definition of conditional probability to compute $P(sex = male | wealth = poor)$:

$$P(sex = male | wealth = poor) = 0.46/0.75 = 0.61$$

Class Exercise. Compute $P(wealth = poor | sex = male)$.

3.4 Classification using the joint probability distribution

At last! It's time to return to the task of classification.

Let the attributes of the instances be $A_1, \ldots, A_n$. The new instance, which is to be classified, will be described by a set of attribute-value pairs, $\{A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n\}$. Let the set of class labels be L.

We can use the joint probability distribution to classify a new instance as follows:

1. For each class label $cl \in L$, use the joint probability distribution to compute the conditional probability for that class given the description of the new instance:

   $$P(class = cl | A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)$$

   (To emphasise: you are computing this for each $cl \in L$, so the number of calculations you perform is $|L|$, the number of class labels.)

2. Return the class label with the highest probability.

For example, suppose we wish to classify a person whose descriptive attributes are $\{sex = male, hours\_worked = 40.5+\}$. Does this person belong to the *poor* or *rich* class?

1.

$$\begin{aligned} P(wealth = poor | sex = male, hours\_worked = 40.5+) &= \frac{P(wealth = poor, sex = male, hours\_worked = 40.5+)}{P(sex = male, hours\_worked = 40.5+)} \\ &= \frac{0.13}{0.13 + 0.11} \\ &= 0.54 \end{aligned}$$

$$\begin{aligned} P(wealth = rich|sex = male, hours\_worked = 40.5+) &= \frac{P(wealth = rich, sex = male, hours\_worked = 40.5+)}{P(sex = male, hours\_worked = 40.5+)} \\ &= \frac{0.11}{0.13 + 0.11} \\ &= 0.46 \end{aligned}$$

2. Return the class with the highest conditional probability. In this case, the person is predicted to be *poor*.

Note that you did not really need to compute the divisor (i.e. the expression below the line)! For each class label whose conditional probability you compute, the divisor will be the same. So only the dividend (above the line) determines which class wins.

3.5 The bad news

This approach to inference in general and classification in particular does not scale well. In realistic problems there will be many more variables, and some will have many more values. The table will be too big. No human could estimate all the probabilities. Equally, no dataset will provide enough data to reliably estimate the probabilities based on frequency in the dataset.