Course on Artificial Intelligence and Intelligent Systems

Examples and exercises for conditional probabilities and Bayesian reasoning

Henning Christiansen
Roskilde University, Computer Science Dept.
 Version 14 oct 2005

1 Motivation and overview

This note refers to part chapter 3 of the course textbook by M. Negnevitsky. We find the text very problematic, and we recommend our reader to skip section 3.3 and onwards.

In this note we provide firstly examples to motivate and explain the standard definitions reviewed in section 3.2 (the textbook lacks examples completely). Secondly, we give a new account on Bayesian reasoning which we, for clarity, release from the terminology of expert systems, and we provide informative examples.

2 Comments and examples to standard notions from probability theory

In this section, we refer specifically to the formulas of section 3.2 in the textbook, and you are supposed to have the book opened in front of you while you read the following.

The purpose of the mentioned section of the textbook is to introduce standard notions of probability theory such as probability for random variables and conditional probabilities. However, no motivating examples are given, so unless you are familiar with probability theory already, you are most likely to have difficulties in understanding the text.

In this note, we provide two related examples that you may use for your reading, and we will go through the detailed calculations at the course (as exercises).

First, we will correct one unclear point in the text. Middle p. 60 reads "The concept of conditional probability introduced so far considered that event A was dependent on event B." In fact, all results and notions referred to goes also for the special case of independent random variables.

A note on terminology: A *random variable* means a variable whose value is determined by an experiment, whose outcome is determined by a probability distribution. If case such a variable V may take a value in a set, say $\{$*head*, *tail*$\}$, an *event* refers to the case that V takes a specific value. So one event may be the observation that an experiment resulted in $V =$ *head*; the textbook uses only the term "event".

You should also be aware that the textbook's *definition* of probability based on a ratio between numbers of "successes" and "number of possible outcomes" is very imprecise and should only be taken as an intuitive characterization; as a definition it is nonsense. However, the so-called *law of large numbers* indicates, with suitable definitions, that when repeating the experiment n times, the indicated ratio will converge to the probability as $n \to \infty$.

We will give two examples to illustrate conditional probability, the first one which, for simplicity, considers independent variables. In a second example, we use dependent random variables.

Both examples concern the colour of peoples' hair, and we will make a few simplifying assumptions.

- The colour of a person's hair is uniquely defined.
- This colour can be one of blond, red, black, brown, and if the person has no hair, we say for uniformity that the colour is "bald".

2.1 First example with independent events, plus comments to the formulas

We consider here a population of young people, which means that there are no bald persons, and the distribution of hair colour is assumed to be the same for boys and girls.

Consider the experiment of selecting, at random, a person from this population, and we assume two random variables, one for the sex of the chosen person and another for the person's colour of hair. Referring to the formulas in the textbook, p. 59–61 (middle), we assume for now, the following types of events:

- A means that the chosen person is a girl.
- B means that the chosen person has red hair.

We assume a probability of 0.5 for A (independently of the person's colour of hair) and a probability of 0.1 for B (independent of whether the person is a boy or a girl).

As an illustration, we consider a subset of the population consisting of 100 persons whose characteristics correspond perfectly to the indicated probabilities, i.e.,

- 50 are girls of which 5 have red hair,
- 50 are boys of which 5 have red hair.

We refer to this sub-population as a *perfectly representative selection*. Conditional probability $P(A|B)$ means, thus, the probability that the chosen person in the experiments is a girl, provided that we know that the person has red hair. As we know this should be 0.5, and we notice in the perfectly representative selection that exactly 5 out of the 10 red-haired persons are girls. Compare this argument with the pseudo-math formula (3.6).

The event $A \cap B$ means that both A and B are case as the outcome of an experiment, i.e., a girl with red hair was selected. We would expect $P(A \cap B)$ to be $0.5 \times 0.1 = 0.05$ which is consistent with the ratio in the perfectly representative selection, 5 red-haired girls. In fact, the standard definition in probability theory defines two random variables X and Y to be *independent*, if the obey the following law.

$$P(X = a \ \cap \ Y = b) = P(X = a) \times P(Y = b)$$

for any values a and b that X, resp., Y may take. Here $X = a$ denotes the event that random variable X takes value a. It follows, furthermore, from standard definitions of probability theory, for independent random variables X and Y, that

$$P(X = a|Y = b) = P(X = a) \qquad \text{and} \qquad P(Y = b|X = a) = P(Y = b).$$

Exercise

Insert our current definitions of A and B in formula (3.7), and verify that it is correct for this example. Provide an intuitive reading of each subexpression in the equation, and verify that the analogous ratios in the perfectly representative selection satisfy similar equations.

Exercise

Do exactly the same for formulas (3.8), (3.9), (3.10), and (3.11).

To illustrate the meaning of formulas (3.12–3.14), let us for $B_1, ..., B_n$ assume $n = 4$ and name the events corresponding to different colours of hair; we assume still the population of young people with independence between sex and colour of hair. The perfectly representative selection conform with the probabilities, and we assume furthermore

- $P(\mathsf{blond}) = 0.4$, thus 40 blond persons in the perfectly repr. sel. (20 girls and 20 boys),
- $P(\mathsf{brown}) = 0.3$, thus 30 brown-haired persons in the perfectly repr. sel. (15 girls and 15 boys),
- $P(\mathsf{black}) = 0.2$, thus 20 black-haired persons in the perfectly repr. sel. (10 girls and 10 boys),
- $P(\mathsf{red}) = 0.1$, thus 10 red-haired persons in the perfectly repr. sel. (5 girls and 5 boys).

(So "red" is the event that we so far called B). We notice, by our initial assumptions, that these four events satisfy the condition of being mutually exclusive. A formal definition of *"mutually exclusive"* could state that $P(\mathsf{blond} \cap \mathsf{black}) = 0$ and $P(\mathsf{blond} \mid \mathsf{black}) = 0$ and similarly for any other combination of different events chosen among $\{\mathsf{blond}, \mathsf{brown}, \mathsf{black}, \mathsf{red}\}$.

Exercise

Check formula (3.12) by inserting our current definitions for events, A still meaning "girl", and with $\{\mathsf{blond}, \mathsf{brown}, \mathsf{black}, \mathsf{red}\}$ for $B_1, \ldots, B_n$. Provide an intuitive reading of each subexpression in the equation, and verify that the analogous ratios in the perfectly representative selection satisfy similar equations.

Equations (3.13–14) refer to the case where $B_1, \ldots, B_n$ satisfies a property referred to as *"exhaustive"*. The meaning of this can be written as a formula

$$P(B_1 \cup \cdots \cup B_n) = 1$$

where $\cup$ can be read as "or". According to our assumptions, $\{\mathsf{blond}, \mathsf{brown}, \mathsf{black}, \mathsf{red}\}$ is exhaustive:

$$P(\mathsf{blond} \cup \mathsf{brown} \cup \mathsf{black} \cup \mathsf{red}) = 1$$

Furthermore, since they are exhaustive *and* independent, we have also the following formula.

$$P(\mathsf{blond}) + P(\mathsf{brown}) + P(\mathsf{black}) + P(\mathsf{red}) = 1$$

As a little aside, let us consider cases when this sum does not equal to one.

- When the events are independent, but not exhaustive, the sum may be < 1 (intuitively because some samples are not counted, e.g., the green-haired).
- When the events are exhaustive, but not independent, the sum may be > 1 (intuitively because some samples are counted more that once, e.g., some one with black stripes in the brown hair).

Wrong beliefs about independence or exhaustiveness are typical sources of errors in probabilistic models.

Exercise

The step from formula (3.12) to (3.13) may appear a bit obscure as no explanation is given. Check the relationship between the two formulas for our current example, A meaning "girl", and with $\{\mathsf{blond}, \mathsf{brown}, \mathsf{black}, \mathsf{red}\}$ for $B_1, \ldots, B_n$.

We will come back to formulas (3.15–3.17) which becomes more interesting when we consider events that are not necessarily independent. ***Notice that we so far never used the assumption that colour of hair and sex was independent.***

2.2 Second example, dependent events

Bayes' formula and its various consequences find their most interesting applications for cases when the events combined in conditional probabilities are dependent. The definition of "dependent" is simple: Two random variables or, alternatively, two events are *dependent* if they are not independent. This formulation is not a joke, but it has the precise meaning that we cannot trust laws anymore such as $P(A|B) = P(A)$ and $P(A \cap B) = P(A) \times P(B)$ when A and B are dependent.

Instead of the population of young people that we referred to above, we consider now a population of grown-ups, where we still consider colour of hair (or lack of hair). Here we assume a larger variation in hair colour, which we can say is due to the fact that some men gets bald (and very few women), and that women more frequently goes to the hairdresser and have their colour of hair changed; furthermore, the nightlife is so boring that some of the younger men have left for another country (and why the younger women stay, no one knows). Clearly, colour of hair depends now on sex in some way or another — and vice versa.

Instead of giving probabilities directly, we describe instead the frequencies for a perfectly representative selection of 100 persons — perfect in the sense that they collectively respect the probabilities for the mentioned events.

Event A \ Event B_i	blond	brown	black	red	blue	bald	(sum)
woman	30	5	15	2	7	1	60
man	12	8	5	4	1	10	40
(sum)	42	13	20	6	8	11	100

Exercise

Write down (as numbers between 0 and 1) all probabilities and conditional probabilities, e.g., $P(\mathsf{woman}), P(\mathsf{blond}), P(\mathsf{brown}|\mathsf{man}), P(\mathsf{man}|\mathsf{red})$, etc., that you can derive from this table.

Exercise

Check the relationship between the two formulas (3.12) to (3.13) for our current example in two versions, A meaning "woman" in the first and "man" in the second, and in both cases $\{\mathsf{blond}, \mathsf{brown}, \mathsf{black}, \mathsf{red}, \mathsf{blue}, \mathsf{bald}\}$ for $B_1, \ldots, B_n$.

3 Bayesian reasoning, the principle

Let us consider again the example of section 2.2 and formula (3.17) from the text book. We let $A = \mathsf{woman}$ and $B = \mathsf{red}$; notice that $\neg\mathsf{woman} = \mathsf{man}$. With this, formula (3.17) becomes the following.

$$P(\mathsf{woman}|\mathsf{red}) = \frac{P(\mathsf{red}|\mathsf{woman}) \times P(\mathsf{woman})}{P(\mathsf{red}|\mathsf{woman}) \times P(\mathsf{woman}) + P(\mathsf{red}|\mathsf{man}) \times P(\mathsf{man})} \tag{1}$$

Assume, now, for some strange reason, that we do not have access to conditional probabilities of the kind $P(\mathit{sex}|\mathit{colour})$ but we know other conditional probabilities $P(\mathit{colour}|\mathit{sex})$ plus the probabilities $P(\mathsf{woman})$ and $P(\mathsf{man})$. In this case we can use the formula above to determine the probability that an observed red-haired person happens to be a woman.

Such a judgement might be interesting in case a red-haired person (dressed in a big coat, but not wearing a hat), has been seen running away from the scene of a crime, and the police has two usual suspects in custody, both red-haired, one woman and one man.

Well, how can it be the case that we know, say $P(\mathsf{red}|\mathsf{woman})$ but not $P(\mathsf{woman}|\mathsf{red})$? In section 2.2 above, we assumed a perfectly representative selection given and that we knew all properties of all individuals, so that all probabilities, including all sorts of conditional ones, could be found by simple counting.

In reality, we very seldom have a perfectly representative selection at hand, unless we have all information available for every person in the society. In general it may be difficult to have sufficient data for estimating all probabilities.

More importantly, it may be the case that we have some prior knowledge which in a particular situation differs from the perfectly representative selection. In our example, we may know that usual suspects typically are men, and in this case we might expect that a probabilistic model should be able to lower the probability that the observed read-haired be a woman.

Let us make some important concepts clear. The unconditional probabilities $P(\mathsf{woman})$ and $P(\mathsf{man})$ are called *prior probabilities* as the represents assumptions about the overall situation before any observation of an event is made. The calculated conditional probability, in this example $P(\mathsf{woman}|\mathsf{red})$, is called a *posterior probability*, as it is a probability calculated after a specific event has been observed (e.g. "red"). If, in our example, we assume that the person who actually performed the crime is taken at random from the population concerned with sex, this represents prior probabilities $P(\mathsf{woman}) = 0.6$ and $P(\mathsf{man}) = 0.4$. However, if we know that 80% of all criminals are men, we may use instead in our model, the prior probabilities $P(\mathsf{woman}) = 0.2$ and $P(\mathsf{man}) = 0.8$.

With Bayes' formula (in the appearance (1) above) we can predict other conditional probabilities concerning the sex of the observed read-haired person, taking into account that that this person is not any person but a criminal. Notice that the shift of prior probabilities does not indicate a shift in the conditional probabilities for hair colours (no assumption is made that, say, red-haired men are more or less criminal than others).

Exercise

Calculate the conditional probability $P(\mathsf{woman}|\mathsf{red})$ for the two different prior probability distributions, $P(\mathsf{woman}) = 0.6$, $P(\mathsf{man}) = 0.4$ vs. $P(\mathsf{woman}) = 0.2$ and $P(\mathsf{man}) = 0.8$.

In the example above where the task is to identify the most likely person who performed some crime, it may be an ethical question whether the prejudice expressed in the different prior probabilities should be taken into account.

We will consider another example in which it seems more obvious to consider different prior probabilities. We consider a certain decease for which you can go to the hospital and get a test; however, the test is not perfect so the test will indicate so-called false positives and false negatives.

So let us replace "woman" by "decease" (for "having the decease") and "red" by "pos" (for "test is positive"). Formula (1) now becomes the following

$$P(\mathsf{decease}|\mathsf{pos}) = \frac{P(\mathsf{pos}|\mathsf{decease}) \times P(\mathsf{decease})}{P(\mathsf{pos}|\mathsf{decease}) \times P(\mathsf{decease}) + P(\mathsf{pos}|\mathsf{no\text{-}decease}) \times P(\mathsf{no\text{-}decease})} \qquad (2)$$

If we assume that the decease is fatal or in other ways develops to become apparent during the lifetime of the patient, we can estimate the probability that a randomly chosen person has (or has not) the decease from statistics about the entire population.

For the given test method, we assume also that a series of experiments have been made in order to gather statistics which makes it possible to judge its precision in terms of true/false positives/negatives, including the probabilities $P(\mathsf{pos}|\mathsf{decease})$ and $P(\mathsf{pos}|\mathsf{no\text{-}decease})$, that we can use in our formula.

Exercise

Let $P(\mathsf{pos}|\mathsf{decease}) = 0.8$ and $P(\mathsf{pos}|\mathsf{no\text{-}decease}) = 0.1$. Consider the case that a person goes to the test as part of a general health check without any specific suspicion that he or she should be more or less exposed to the decease than any other people. Assume prior probabilities based on statistics for the entire population saying that 5% has this decease, i.e., $P(\mathsf{decease}) = 0.05, P(\mathsf{no\text{-}decease}) = 0.95$. Calculate the posterior probability that our sample person has the decease provided that he has been tested positive.

Exercise

The day after, another person comes for the test, who is working at an nuclear power plant. It is known that of all nuclear power plant workers, 60% of them have the decease. In case this guy is tested positive, you should calculate his probability to have the decease, given the indicated posterior probabilities.

Finally, a colleague of the last mentioned goes to the test and is tested negative. What is his probability to have the decease.

4 An example of Bayesian reasoning, Bayesian networks

Bayesian theory gives also rise to methods for general classification, and identification of spam mails is a standard example.

We assume here that a number, say n, of indicators of spam mails are being considered. For political correctness, we consider some unknown language in which quack, wack, boink, etc. are words that occur more frequently in spam mails that other mails.

The theory we have developed so far is not very well suited for the case with multiple indicators, may or may not be observed. Assume that we consider a mail which contains the three word quack, wack, boink, and no other problematic words. In that case we would like to have to conditional probabilities

$$P(\mathsf{quack} \cap \mathsf{wack} \cap \mathsf{boink} \mid \mathsf{spam}) \qquad \text{and} \qquad P(\mathsf{quack} \cap \mathsf{wack} \cap \mathsf{boink} \mid \cap\, \mathsf{no\text{-}spam})$$

available. In principle, we could gather sufficient amount of statistics to get an estimate, but in the general case with n indicators there are 2^n different combinations of present/not-present for which conditional probabilities should estimated. Without giving detailed arguments, it appears that the amount of statistics, which should be based on manually classified mails, needs to be enormous.

An extension of the method we have seen so far, which is better for such applications, is so-called Bayesian networks. We may interested readers refer to the following book (which is just one out of many).

R.E. Neapolitan: *Learning Bayesian Networks.* Prentice Hall, 2004.