Lecture 14: Bayes formula

Conditional probability has many important applications and is the basis of Bayesian approach to probability:

- Consider events $B_1, B_2, \cdots B_n$ which are pairwise disjoint and

  $$B_1 \cup B_2 \cup \cdots \cup B_n$$

  These events are called the *hypotheses*. The probabilities

  $$P(B_1), \quad P(B_2), \quad \cdots \quad P(B_n)$$

  are called the *prior probabilities*. One may think of these probabilities as describing our a-prior level of knowledge (or ignorance) about the the experiment: $P(B_i)$ gives us the probability that our hypothesis B_i is correct.

- We consider an event E which gives us information about which hypothesis is correct. The event E is called the *evidence*. The information about the the hypotheses is given through the conditional probabilities

  $$P(E|B_1), \quad P(E|B_2), \quad \cdots \quad P(E|B_n)\,.$$

  which we assume to be known.

- Finally we want to understand how the evidence has informed us about the probability of the hypotheses, i.e. we want to compute

  $$P(B_1|E) \quad P(B_2|E), \quad \cdots \quad P(B_n|E)$$

  These probabilities are called the *posterior probabilities*.

  To compute $P(B_i|E)$ we recall that

  $$P(B_i \cap E) \;=\; P(E|B_i)P(B_i)$$

  and that

  $$P(E) \;=\; P(E|B_1)P(B_1) + P(E|B_2)P(B_2) + \cdots P(E|B_n)P(B_n)$$

  and so we have

  $$\begin{aligned} P(B_i|E) &= \frac{P(B_i \cap E)}{P(E)} \\ &= \frac{P(E|B_i)P(B_i)}{P(E|B_1)P(B_1) + P(E|B_2)P(B_2) + \cdots P(E|B_n)P(B_n)} \end{aligned}$$

  This formula expresses the posterior probabilities in terms of the prior probabilities and the probabilities of the evidence. They show us how to update our knowledge.

Bayes formula $$P(B_i|E) = \frac{P(E|B_i)P(B_i)}{P(E)} = \frac{P(E|B_i)P(B_i)}{\sum_{k=1}^{n} P(E|B_k)P(B_k)}$$

Example: Suppose Math 478 has two section. In section I there is 12 female and 18 male students. In section II there are 20 female and 15 male students. The professor picks a section at random and then picks a student at random in that section. Compute

1. Probability that the student chosen is a female.
2. The conditional probability that the student is in section II given that she is a female

1. Let us compute first the probability that the student chosen is a female. To do this we condition on whether the professor choose a student from section I or from section II. We have

$$P(\text{section I}) = P(\text{section II}) = \frac{1}{2}$$

and we know the conditional probabilities

$$P(\text{female } |\text{section I}) = \frac{12}{30}$$

$$P(\text{female } |\text{section II}) = \frac{20}{35}$$

So we have

$$\begin{aligned} P(\text{female}) &= P(\text{female } |\text{section I})P(\text{section I}) + P(\text{female } |\text{section II})P(\text{section II}) \\ &= \frac{12}{30} \times \frac{1}{2} + \frac{20}{35} \times \frac{1}{2} = \frac{17}{35} \end{aligned}$$

2. To compute $P(\text{section } II \mid \text{ female })$ we use Bayes formula

$$P(\text{section } II| \text{ female }) = \frac{P(\text{ female } |\text{section } II)P(\text{section II})}{P(\text{female})}$$

Using the result in part 1. we find

$$P(\text{section } II| \text{ female }) = \frac{\frac{20}{35} \times \frac{1}{2}}{\frac{12}{30} \times \frac{1}{2} + \frac{20}{35} \times \frac{1}{2}} = \frac{10}{17}$$

Example: False positives, Sensitivity and Specificity. Bayes formula is very much used in epidemiology. Suppose we deal with a disease and we have test for the disease. We know

- The *sensitivity* of the test is 99%, this means that if you have the disease, the test is positive with probability 0.99.
- The *specificity* of the test is 98%, this means that if you do not have the disease, the test is negative with probability 0.98.
- The *prevalence* of the disease is one in two hundred, this means that the probability to carry the disease is 0.005.

In the Bayesian framework, the hypotheses are the events D to have the disease and the events $\overline{D}$ of not having the disease and the prior probabilities are

$$P(D) \,=\, 0.005\,, \quad P(\overline{D}) = .995\,.$$

The evidence is the test being positive (+) or negative (-). The sensitivity and specificity give the conditional probabilities

$$P(+|D) = 0.99\,, \quad P(-|D) = 0.01\,, \quad P(+|\overline{D}) = 0.02\,, \quad P(-|\overline{D}) = 0.98$$

The posterior probabilities are given by Bayes formula

$$P(D|+) \,=\, \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|\overline{D})P(\overline{D})} \,=\, \frac{.99 \times .005}{.99 \times .005 + .02 \times .995} \,=\, 0.1991$$

and

$$P(D|-) \,=\, \frac{P(-|D)P(D)}{P(-|D)P(D) + P(-|\overline{D})P(\overline{D})} \,=\, \frac{.01 \times .005}{.01 \times .005 + .98 \times .995} \,=\, 0.00005$$

The quantity $P(D|+)$ is the probability that someone with a positive test actually has the disease. It is called in epidemiology the *positive predictive value* of a test. The quantity $P(\overline{D}|-)$ is the probability that someone with a negative test actually does not have the disease. It is called in epidemiology the *negative predictive value* of a test.

In the example above the predictive value of the test is fairly poor: only around 20% percent of the people testing positive have the disease. This can be easily understood: if the disease is fairly rare then the false positive will be much more numerous than the true positive. On the other hand the negative predictive value is excellent.

Exercise 1: A certain type of cancer is known to in 2 percent of the population of the males in their fifties. A test for the disease has been is advertised by a pharmaceutical company to have %3 of false negative and %1 of false positive.

1. Compute the probability that you have cancer if you are tested positive.

2. To make sure that you really have cancer an invasive and expensive surgery is needed. Your health insurance company is not willing to pay for this unless the pharmaceutical company improves its test in such way that that at least %90 of people who are tested positive actually have the disease. How low should be the rate of false positive for the test to reach this goal? (Assume that the rate of false negative remains the same).

Exercise 2: To do a certain commuting travel in the city of B., the probability to be delayed more than half an hour is 0.57 if you travel by car, 0.28 if you take the bus (which can use an HOV traffic lane), and 0.05 if you take a commuter train. Company Green wishes to encourages his employees to use public transportation as well as minimize lost hours and so it provides a financial incentive to his employees to use bus or train. Bob is late this morning, his supervisor would like to find out how likely it is that Bob actually did not use his car. To do this assign prior probabilities that Bob takes his car, bus, or train and use then Bayes formula. Discuss how you should assign prior probabilities.