CS 234 Data, Analytics, and Visualization (Fall 2017)

09/12/2017 Class Worksheet Nr. 1

The Oreo Cookie Problem

[slightly modified from an example by Allen Downey]

Suppose there are two bags or Oreo cookies.

Bag_1 contains 30 vanilla and 10 chocolate oreos.
Bag_2 contains 20 vanilla and 20 chocolate oreos.

The two bags are indistinguishable. Suppose we choose one bag at random. Then from this bag, without looking, choose one cookie at random.

We observe that the chosen cookie is Vanilla.

Question: What is the probability that the cookie came from Bag_1?

We will write this probability as P(Bag_1 | Vanilla). That is: the conditional probability of having chosen Bag_1, given that we know that the oreo is Vanilla.

Notice that this probability is different from that of the inverse event: given that we know we chose Bag_1, what is the probability of drawing a vanilla oreo: P(Vanilla | Bag_1). We can easily calculate this probability:

P(Vanilla | Bag_1) = _____

In general, P (A|B) is different from P(B|A) and there is no obvious way to calculate P(Bag_1 | Vanilla) from P(Vanilla | Bag_1). Here is where Bayes' Theorem comes to rescue!

Baeys's Theorem

Let's derive the theorem together. The probability of two joint events P(A ^ B) is calculated as:

P(A ^ B) = P(A) * P(B | A) [1]

Is there an alternative way we can write this statement, knowing that A and B can be interchangeable?

_____ [2]

Let's combine [1] and [2]:

_____ and then derive the formula for P(A|B)

P(A|B) = _____ [Bayes's Theorem]

Exercise 1: Let's calculate the P(Bag_1 | Vanilla) by using Bayes's Theorem.
To simplify notation, let's denote Bag_1 as B1 and Vanilla as V. Initially, let's fill in the Bayes' theorem formula using our notation:

P(B1|V) = _____

Calculate below each of the components of the formula:

P(B1) =

P(V|B1) = [look it up in the previous page]

P(V) = (over all bags of oreo cookies)

Final value: P(B1|V) =

Exercise 2: What is the probability **P(Bag_2 | Vanilla)**?

Updating probability in light of new evidence (or data)

Let's denote H (hypothesis) and D (data or evidence):

$$p(H|D) = \frac{p(H)\ p(D|H)}{p(D)}$$

P(H) → is the probability of the hypothesis before we see the data, called the prior probability, or just **prior**.
P(H|D) → is what we want to compute, the probability of the hypothesis after we see the data, called the **posterior**.
P(D|H) → is the probability of the data under the hypothesis, called the **likelihood**.
P(D) → is the probability of the data under any hypothesis, called the **normalizing constant**

The prior might be given or we make a guess based on our beliefs. The likelihood is the easiest to calculate (if we know the bag, we can calculate the cookie probability). The normalizing constant can be tricky. We make some simplifying assumptions. We imagine that we have a set of hypotheses that are:

(a) Mutually exclusive: at most one hypothesis can be true
(b) Collectively exhaustive: there are no other possibilities, at least one hypothesis is true

If these assumptions hold, we can use the law of total probability. In case of two hypotheses, we'll write:

P(D) = P(H1)*P(D|H1) + P(H2)*P(D|H2)

Your Turn: Suppose that you are worried that you might have a rare disease. You decide to get tested, and suppose that the testing methods for this disease are correct 99 percent of the time (in other words, if you have the disease, it shows that you do with 99 percent probability, and if you don't have the disease, it shows that you do not with 99 percent probability). Suppose this disease is actually quite rare, occurring randomly in the general population in only one of every 10,000 people.
If your test results come back positive, what are your chances that you actually have the disease?
Do you think it is approximately: (a) .99, (b) .90, (c) .10, or (d) .01?