Data Analysis 2: Foundations of Statistics

Instructor: Arieda Muço, Fall 2017

Bayes' Theorem

- Is important in many applications in Data Analysis, Machine Learning, and Economics (in general)
- Describes the probability of an event, based on the conditions that might be related to that event.
- It allows us to use previously known information to access the likelihood of another event.

Bayes' Theorem Display

- We usually display the theorem in one of the two following equations.

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} = \frac{p(B|A)p(A)}{p(B|A)p(A)+p(B|\sim A)p(\sim A)}$$

- Joint probability is the probability that two things are contemporaneously true.

$$p(A\&B) = p(B\&A)$$

$$p(A\&B) = p(A)p(B|A)$$

$$p(B\&A) = p(B)p(A|B)$$

$$\Downarrow$$

$$p(B|A) = \frac{p(B)p(A|B)}{p(A)}$$

- The alternative way to write the formula comes from the law of total probabilities.

  - Where $p(A)$ and $p(\sim A)$

Are two mutually exclusive events

The cookie problem!

Two bowls of cookies. Bowl 1 contains 30 vanilla and 10 chocolate cookies

Bowl 2 contains 20 vanilla and 20 chocolate cookies

Select a cookie at random without looking and the cookie is vanilla. What is the probability that it came from Bowl 1?

The cookie problem!

We want:

$$p(\text{bowl 1}|\mathit{vanilla})$$

What do we know?

$$p(\mathit{vanilla}|\text{bowl 1}) = \frac{3}{4}$$

$$p(\text{bowl 1}) = \frac{1}{2}$$

$$p(\text{vanilla}) = \frac{5}{8}$$

Event changes as we see the data

- Another way to think about the Bayes' theorem is to update the probability of hypothesis, H, as we see the data D

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

$p(H)$ **Prior** probability

$p(H|D)$ is the probability of the hypothesis after we have seen the data, **posterior**

$p(D|H)$ **Likelihood**

$p(D)$ probability of the data under any hypothesis

$$p(D) = ?$$

- In some cases we can compute it using the law of total probability, which says that if there are two exclusive ways that something might happen, you can add up probabilities as follows:

$$p(D) = p(\text{bowl } 1)p(D|\text{bowl } 1) + p(\text{bowl } 2)p(D|\text{bowl } 2)$$

$$p(D) = (1/2)(3/4) + (1/2)(1/2)$$