PRINCIPLES OF DATA SCIENCE

2021 - 2022

Assoc. Prof. Dr. Salha Alzahrani

Department of Computer Science

جامعة الطائف
TAIF UNIVERSITY

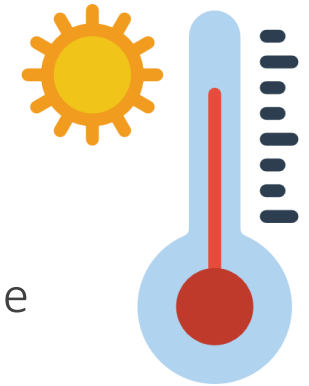
CHAPTER 6: ADVANCED PROBABILITY

- COLLECTIVELY EXHAUSTIVE EVENTS
- BAYESIAN IDEAS REVISITED
- BAYES THEOREM
  - Example – Titanic
  - Example – medical studies
- RANDOM VARIABLES
  - Discrete random variables

COLLECTIVELY EXHAUSTIVE EVENTS

Collectively exhaustive events

- When given a set of two or more events, if at least one of the events must occur, then such a set of events is said to be collectively exhaustive.
- Consider the following examples:
  - Given a set of events {temperature < 60, temperature > 90}, these events are not collectively exhaustive because there is a third option that is not given in this set of events: The temperature could be between 60 and 90. However, they are mutually exhaustive because both cannot happen at the same time.



  - In a dice roll, the set of events of rolling a {1, 2, 3, 4, 5, or 6} are collectively exhaustive because these are the only possible events, and at least one of them must happen.

BAYESIAN IDEAS REVISITED

Bayesian ideas revisited

- When speaking about Bayes, you are speaking about the following three things and how they all interact with each other:
  - ✓ A prior distribution
  - ✓ A posterior distribution
  - ✓ A likelihood

- Basically, we are concerned with finding the posterior.

- Another way to phrase the Bayesian way of thinking is that data shapes and updates our belief. We have a prior probability, or what we naively think about a hypothesis, and then we have a posterior probability, which is what we think about a hypothesis, given some data.

BAYES THEOREM

Bayes theorem

- Bayes theorem is the big result of Bayesian inference. Recall the following:
  - P(A) = The probability that event A occurs
  - P(A|B) = The probability that A occurs, given that B occurred
  - P(A, B) = The probability that A and B occurs
  - P(A, B) = P(A) * P(B|A)

- That last bullet can be read as the probability that A and B occur is the probability that A occurs times the probability that B occurred, given that A already occurred.

Bayes theorem

- We know that:
  P(A, B) = P(A) * P(B|A)
  P(B, A) = P(B) * P(A|B)
  P(A, B) = P(B, A)
- So: *P(B) * P(A|B) = P(A) * P(B|A)*
- Dividing both sides by P(B) gives us Bayes theorem, as shown:

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}$$

- You can think of Bayes theorem as follows:
  - It is a way to get from P(A|B) to P(B|A) (if you only have one)
  - It is a way to get P(A|B) if you already know P(A) (without knowing B)

Bayes theorem

- Let's try thinking about Bayes using the terms hypothesis and data.
- Suppose H = your hypothesis about the given data and D = the data that you are given.
- Bayes can be interpreted as trying to figure out P(H|D) (the probability that our hypothesis is correct, given the data at hand).

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

  - P(H) is the probability of the hypothesis before we observe the data, called the prior probability
  - P(H|D) is what we want to compute, the probability of the hypothesis after we observe the data, called the posterior
  - P(D|H) is the probability of the data under the given hypothesis, called the likelihood
  - P(D) is the probability of the data under any hypothesis, called the normalizing constant

EXAMPLE – TITANIC

Example – Titanic

- A very famous dataset involves looking at the survivors of the sinking of the Titanic in 1912. We will use an application of probability in order to figure out if there were any demographic features that showed a relationship to passenger survival. Mainly, we are curious to see if we can isolate any features of our dataset that can tell us more about the types of people who were likely to survive this disaster.

```
titanic = pd.read_csv(data/titanic.csv')#read in a csv
titanic = titanic[['Sex', 'Survived']]  #the Sex and Survived column
titanic.head()
```

	Sex	Survived
0	male	no
1	female	yes
2	female	yes
3	female	yes
4	male	no

Example – Titanic

- In the preceding table, each row represents a single passenger on the ship, and, for now, we are looking at two specific features: the sex of the individual and whether or not they survived the sinking. For example, the first row represents a man who did not survive while the fourth row represents a female who did survive.
- Let's start by calculating the probability that any given person on the ship survived, regardless of their gender. To do this, let's count the number of yeses in the Survived column and divide this figure by the total number of rows, as shown here:

```
num_rows = float(titanic.shape[0]) # == 891 rows
p_survived = (titanic.Survived=="yes").sum() / num_rows # == .38
p_notsurvived = 1 - p_survived                          # == .61
```

- Note that I only had to calculate P(Survived), and I used the law of conjugate probabilities to calculate P(Died) because those two events are complementary.

Example – Titanic

- Now, let's calculate the probability that any single passenger is male or female:

```
p_male = (titanic.Sex=="male").sum() / num_rows     # == .65
p_female = 1 - p_male # == .35
```

- Now let's ask ourselves a question, did having a certain gender affect the survival rate?
- P(Survived|Female) or the chance that someone survived given that they were a female. For this, we need to divide the number of women who survived by the total number of women, as shown here:

$$P(Survived \mid Female) = \frac{P(Female\ AND\ Survived)}{P(Female)}$$

```
number_of_women = titanic[titanic.Sex=='female'].shape[0] # == 314
women_who_lived = titanic[(titanic.Sex=='female') & (titanic.
Survived=='yes')].shape[0]                          # == 233
p_survived_given_woman = women_who_lived / float(number_of_women)
p_survived_given_woman            # == .74
```

- That's a pretty big difference. It seems that gender plays a big part in this dataset.

EXAMPLE – MEDICAL STUDIES

Example – medical studies

- A classic use of Bayes theorem is the interpretation of medical trials.
- Routine testing for illegal drug use is increasingly common in workplaces and schools. The companies that perform these tests maintain that the tests have a high sensitivity, which means that they are likely to produce a positive result if there are drugs in their system. They claim that these tests are also highly specific, which means that they are likely to yield a negative result if there are no drugs.
- On average, let's assume that the sensitivity of common drug tests is about 60% and the specificity is about 99%. It means that if an employee is using drugs, the test has a 60% chance of being positive, while if an employee is not on drugs, the test has a 99% chance of being negative.
- Now, suppose these tests are applied to a workforce where the actual rate of drug use is 5%.

Example – medical studies

- The real question is of the people who test positive, how many actually use drugs?
- In Bayesian terms, we want to compute the probability of drug use, given a positive test.
  - Let D = the event that drugs are in use
  - Let E = the event that the test is positive
  - Let N = the event that drugs are NOT in use

- We are looking for P(D|E). By using Bayes theorem , we can extrapolate it as follows:

$$P(D|E) = \frac{P(E|D)P(D)}{P(E)}$$

- The prior, P(D) is the probability of drug use before we see the outcome of the test, which is 5%.
- The likelihood, P(E|D), is the probability of a positive test assuming drug use, which is the same thing as the sensitivity of the test.
- The normalizing constant, P(E), is a little bit trickier.

Example – medical studies

- We have to consider two things: P(E and D) as well as P(E and N). Basically, we must assume that the test is capable of being incorrect when the user is not using drugs.

$$P(E) = P(E \text{ and } D) \text{ or } P(E \text{ and } N)$$

$$P(E) = P(D)P(E|D) + P(N)P(E|N)$$

$$P(E) = .05 * .6 + .95 * .01$$

$$P(E) = 0.0395$$

- So, our original equation becomes as follows:

$$P(D|E) = \frac{.6 * .05}{0.0395}$$

$$P(D|E) = .76$$

- This means that of the people who test positive for drug use, about a quarter are innocent!

RANDOM VARIABLES

Random variables

- A random variable uses real numerical values to describe a probabilistic event.
- In our previous work with variables (both in math and programming), we were used to the fact that a variable takes on a certain value.
- For example, we might have a triangle in which we are given a variable h for the hypotenuse, and we must figure out the length of the hypotenuse. We also might have, in Python: x = 5. Both of these variables are equal to one value at a time.
- In a random variable, we are subject to randomness, which means that our variables' values are, well just that, variable! They might take on multiple values depending on the environment. A random variable still, as shown previously, holds a value.
- **The main distinction between variables as we have seen them and a random variable** is the fact that a random variable's value may change depending on the situation.

Random variables

If a random variable can have many values, how do we keep track of them all?

- Each value that a random variable might take on is associated with a percentage.
- For every value that a random variable might take on, there is a single probability that the variable will be this value.
- With a random variable, we can also obtain our probability distribution of a random variable, which gives the variable's possible values and their probabilities.
- We generally use single capital letters to denote random variables. For example, we might have:
  - X = the outcome of a dice roll
  - Y = the revenue earned by a company this year
  - Z = the score of an applicant on an interview coding quiz (0-100%)
- Effectively, **a random variable is a function that maps values from the sample space of an event (the set of all possible outcomes) to a probability value (between 0 and 1).**
- Think about the event as being expressed as the following:

$$f(event) = probability$$

- There are two main types of random variables: discrete and continuous.

s.zahrani@tu.edu.sa