CS146 Assignment 1

Diagnostic assessment

This assignment is designed to help you assess your readiness for CS146. All exercises will be graded. You will be provided with a solution set once all CS146 students have submitted their assignments.

Submit your work as a single PDF file. Please type your answers using Google Docs, LaTeX, Jupyter notebooks, CoCalc, or any other software that allows you to neatly type text and math.

Show your work for all exercises! Do not simply turn in final answers.

1. Rules of probability theory

Identify which of the following statements are correct and which are incorrect. Explain in one sentence why you identified each statement as correct or incorrect.

1. $P(A,B) = P(A \mid B)\, P(B)$
2. $P(A) = P(A \mid B)\, P(B) + P(A \mid not\ B)\, P(not\ B)$
3. $P(A) = P(A \mid B)\, P(B) + P(A \mid C)\, P(C) + P(A \mid D) + P(D)$
4. $P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$
5. $P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$

2. Logarithms and probability distributions

What is the log of the probability density function (pdf) of each of the following probability distributions?

- Normal distribution
- Gamma distribution
- Beta distribution

You can find the pdf in the summary on the right of each of the pages linked to above.

Also, you will see lots of Greek symbols used in the pdfs. Now is a good time to start learning their names and how to write them, if you don't know them already. We will use these symbols a lot. See this image for a table of symbols with pronunciations.

3. Normal distribution

If x is distributed according to the normal distribution with mean μ and standard deviation σ, and if $f(x) = x^3 + 2x + 1$.

1. Calculate the expected value of $f(x)$.
2. Calculate the probability $P(\, f(x) > 1\,)$.
3. Write a Python script to confirm your answer to question 2. Generate a lot of random numbers from a normal distribution with a particular mean and standard deviation. Calculate $f(x)$ for each of these random numbers. How many of them are greater than 1? Does that match the probability you calculated in question 2?

4. Marginal and conditional probabilities

In a country with high unemployment, 25.2% of the working population is considered "young", and of the young working population, 37.7% are unemployed. The unemployment rate for those in the working population who are not young, is lower at 21.5%.

Write down or calculate, as required, all the following marginal and conditional probabilities for the working population in this scenario.

- P(young)
- P(not young)
- P(unemployed)
- P(not unemployed)
- P(unemployed | young)
- P(young | unemployed)
- P(unemployed | not young)
- P(not young | unemployed)

5. Inference

You take a guess and think there is a 33% chance that Olivia, 5 year old girl, will be able to read at Grade 1 level by the time she turns 6. (Kids usually start learning to read later, at school, so this isn't a terrible guess.)

An educational expert tells you that of children who are already able to read at Grade 1 level by the time they turn 6 years old, 65% had training (usually by their parents) in pronouncing simple words and writing some letters, and 35% did not. On the other hand, of children who are not already able to read at Grade 1 level by age 6, only 10% had the same training.

1. After receiving this information, what is your revised estimate of the probability that Olivia will be able to read at that level if you learned that her parents are currently busy doing basic training for reading with her?
2. The above is a very simple statistical model. In a paragraph of 80–120 words outline the structure of the model. What are the variables? How are these variables related?

More practice exercises (optional)

I cannot emphasize enough: You need to do mathematics and programming to learn them properly. Reading is not enough. Thinking carefully through what you read helps a lot. Practicing and applying different techniques and concepts ultimately solidifies your understanding. There is no way around it.

Below are additional practice exercises for you to attempt. These are optional and you can choose to do as many or as few as you want. If you get stuck on any of them, contact your instructor with specific questions via email and during office hours. Just saying "I'm stuck" is not enough — explain what you tried and where you got stuck so your instructor can understand your thinking and where you might have missed something or made a mistake.

6. Calculating probabilities

1. How many students are in your CS146 class (including yourself)?
2. Assuming you know nothing about your classmates —
   a. What is the probability that you were born after all of them?
   b. What is the probability that you were born before all of them?
   c. What is the probability that you were born after at least half of them?

7. Python script

Confirm your solution to Exercise 6 numerically by writing a Python script. They're harder than you think and there's a good chance you got them wrong without realizing it.

Generate a list of uniformly distributed random numbers using the `random.uniform()` function in the `numpy` module.

a. Check whether the first number is greater than all other numbers.
b. Check whether the first number is less than all other numbers.
c. Check whether the first number is greater than at least half of the other numbers.

Do this 1,000,000 times and count how often each of the above three events happen. Use your counts to report the approximate probability for each event. These values should match your answers to questions 6a–c in Exercise 6 above. If they do, congratulations! your intuition for probability problems is well-developed. If your answers do not match, try to figure out where you went wrong and update your solutions to Exercise 6.

8. Logarithms

If $p \in [0,1]$ is a probability, $\log \frac{p}{1-p}$ is known as the *log-odds* or the *logit* of p. This function comes up a lot in data modeling.

1. If an event has 10:1 odds of happening, what is the probability of the event? What is the logit of this probability when calculated using base 10, base 2, and base e?
2. If an event has 100:1 odds of happening, what is the probability of the event? What is the logit of this probability when calculated using base 10, base 2, and base e?
3. Why does it not really matter match which base we use when calculating the logit function, as long as we always use the same base in our calculations?

By the way, the inverse of the logit function is called the logistic function, which also comes up a lot in data modeling.

- The logit function maps from a probabilities, in $[0,1]$, to real numbers, in $(-\infty, \infty)$.
- The logistic function maps from the real numbers to probabilities.

9. Bags and cookies

There are two bags containing 3 types of cookies — chocolate, vanilla, and caramel. The first bag has 1 chocolate cookie and 2 vanilla cookies. The second bag has 1 chocolate cookie and 3 caramel cookies. Without looking inside, you stick your hand in one of the two bags and take out a cookie, which turns out to be chocolate. You proceed to eat the cookie.

1. What is the probability that you picked the bag containing the vanilla cookies?
2. If you now take a cookie out of the other bag, what is the probability that the cookie is chocolate, vanilla, or caramel?