DSC 40A - Homework 08

Due: Friday, March 13, 2020

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope on Friday afternoon at 5:00 p.m.

Problem 1.

Suppose A and B are events in a sample space S with $0 < P(A) < 1$ and $0 < P(B) < 1$.

a) If A is a subset of B, can A and B be independent? Prove your answer.

b) If A is disjoint from B, can A and B be independent? Prove your answer.

c) Give an example of two events A and B that are independent, where the sample space $S = \{a, b, c, d, e, f\}$ and the probability distribution is given in the table below.

a	b	c	d	e	f
$\frac{1}{16}$	$\frac{5}{16}$	$\frac{4}{16}$	$\frac{1}{16}$	$\frac{3}{16}$	$\frac{2}{16}$

Problem 2.

This problem will illustrate that independence and conditional independence are not related, in the sense that neither property implies the other.

Consider the sample space $S = \{1, 2, 3, 4, 5, 6\}$ with associated probability distribution $P(s) = \frac{1}{6}$ for each s in S. You can think of S as representing the possible outcomes of rolling a single die.

For each part below, define events A, B, C in this sample space that satisfy the given requirements. Demonstrate that the requirements are satisfied by computing the appropriate probabilities. Make sure to choose events that are neither impossible nor certain, that is, $0 < P(A), P(B), P(C) < 1$.

a) A and B are independent.
A and B are conditionally independent given C.

b) A and B are independent.
A and B are not conditionally independent given C.

c) A and B are not independent.
A and B are conditionally independent given C.

d) A and B are not independent.
A and B are not conditionally independent given C.

Problem 3.

You arrive to campus in one of three ways: on the bus, in your car, or on foot.

The buses often run late and have long lines, so when you take the bus, you have a 50% chance of being late for class. When you drive, you sometimes hit traffic or have trouble finding a parking spot, so you have a 30% chance of being late for class. When you walk to campus, you only have a 5% of arriving late. One day, you arrive late for your midterm, and your professor wonders how you got to school that day.

a) If your professor assumes that you are equally likely to use all three modes of transportation, what will the professor calculate for the probability that you took the bus on the day of your midterm?

b) If your professor happens to know that you take the bus 20% of the time, drive 20% of the time, and walk 60% of the time, what will the professor calculate for the probability that you took the bus on the day of your midterm?

Problem 4.

You are a junior at UCSD who has decided that now is a good time to start laying out post-graduation plans. You are considering the following three occupations: data scientist, software engineer, and business analyst. To get a sense of the background of individuals in each of these occupations, you reach out to current data scientists, software engineers, and business analysts and ask them the following questions:

- What was your college major (data science, computer science, or mathematics)?
- What is your favorite tool (Python, Excel, or Java)?
- What is your favorite work activity (data analysis, programming, or writing/presenting reports)?

You collect a training data set consisting of 30 total individuals in these occupations. You will use a naive Bayes classifier to build a recommender system that you and your friends can use. That is, given someone's college major, favorite tool, and favorite work activity, you want to determine the occupation he or she is most suited for.

a) Use the training data that follows and a naive Bayes classifier to determine the most likely occupation for a mathematics major who likes Python and data analysis.

b) Use the training data *without major* to classify the same student based only on favorite tool language and favorite work activity. Does the classifier perform differently when we remove this important distinguishing feature?

c) In the training data, each of the three occupations were about equally represented. If our training data instead had an imbalance of occupations, how would we expect this to affect the outcome of our naive Bayes classifier? Does training on imbalanced classes affect how the classifier makes predictions? Explain.

d) In the training data, there were three features, each of which has some association with occupation. If instead our training data had included a fourth feature with no association whatsoever to occupation, such as favorite color, how would we expect this to affect the outcome of our naive Bayes classifier? Does including unrelated features affect how the classifier makes predictions? Explain.

Occupation	Major	Favorite Tool	Favorite Work Activity
data scientist	data science	Python	data analysis
data scientist	data science	Python	data analysis
data scientist	data science	Python	programming
data scientist	data science	Python	programming
data scientist	data science	Python	reports
data scientist	data science	Excel	reports
data scientist	computer science	Python	data analysis
data scientist	computer science	Python	programming
data scientist	mathematics	Python	data analysis
data scientist	mathematics	Excel	reports
software engineer	data science	Python	data analysis
software engineer	data science	Python	programming
software engineer	data science	Java	data analysis
software engineer	computer science	Python	data analysis
software engineer	computer science	Python	programming
software engineer	computer science	Java	data analysis
software engineer	computer science	Java	programming
software engineer	computer science	Java	programming
software engineer	mathematics	Python	reports
software engineer	mathematics	Java	data analysis
software engineer	mathematics	Java	programming
business analyst	data science	Python	programming
business analyst	data science	Excel	data analysis
business analyst	data science	Excel	reports
business analyst	computer science	Java	data analysis
business analyst	mathematics	Python	data analysis
business analyst	mathematics	Python	programming
business analyst	mathematics	Python	reports
business analyst	mathematics	Excel	reports
business analyst	mathematics	Excel	reports