# 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.

**Solution:** If $A$ is a subset of $B$, then $P(B|A) = 1$, but $P(B) < 1$, so $P(B|A) \neq P(B)$, which means $A$ and $B$ are not independent. If one event is a subset of the other, then the events are not independent, which makes sense because knowledge of one event tells you something about the chance of the other event occurring.

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

**Solution:** In this case, $P(B|A) = \dfrac{P(A \cap B)}{P(A)} = 0$ because $P(A \cap B) = 0$ for disjoint events. So $P(B|A) = 0$, but $P(B) > 0$, so $A$ and $B$ are not independent. This makes sense because for disjoint events, knowing that one event happens gives complete information about the other event: it does not happen.

**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.

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

**Solution:** Let $A = \{a, b, c\}$. Then

$$P(A) = \frac{1}{16} + \frac{5}{16} + \frac{4}{16} = \frac{10}{16} = \frac{5}{8}$$

If we define $B = \{b, e\}$, then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{b\})}{P(\{b, e\})} = \frac{\frac{5}{16}}{\frac{8}{16}} = \frac{5}{8}.$$

Since $P(A) = P(A|B)$, we know that $A$ and $B$ are independent.

**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 $C$.

**Solution:**

$$A = \{1, 2, 3\}, B = \{1, 4\}, C = \{1, 2, 4, 5\}$$

We know that $A$ and $B$ are independent because

$$P(A) = \frac{1}{2}$$
$$P(B) = \frac{1}{3}$$
$$P(A \cap B) = P(\{1\}) = \frac{1}{6}$$

so $P(A \cap B) = P(A) * P(B) = \frac{1}{6}$.

We know that $A$ and $B$ are conditionally independent given $C$ because

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(\{1, 2\})}{P(\{1, 2, 4, 5\})} = \frac{1}{2}$$
$$P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{P(\{1, 4\})}{P(\{1, 2, 4, 5\})} = \frac{1}{2}$$
$$P((A \cap B)|C) = \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(\{1\})}{P(\{1, 2, 4, 5\})} = \frac{1}{4}$$

so $P((A \cap B)|C) = P(A|C) * P(B|C) = \frac{1}{4}$.

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

**Solution:**

$$A = \{1, 2, 3\}, B = \{1, 4\}, C = \{1, 2, 4\}$$

We showed in part (a) that these same sets $A$ and $B$ were independent.

We know that $A$ and $B$ are not conditionally independent given $C$ because

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(\{1, 2\})}{P(\{1, 2, 4\})} = \frac{2}{3}$$
$$P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{P(\{1, 4\})}{P(\{1, 2, 4\})} = \frac{2}{3}$$
$$P((A \cap B)|C) = \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(\{1\})}{P(\{1, 2, 4\})} = \frac{1}{3}$$

so $P(A|C) * P(B|C) = \frac{2}{3} * \frac{2}{3} = \frac{4}{9} \neq P((A \cap B)|C)$.

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

**Solution:**

$$A = \{1, 2, 3\}, B = \{1, 2, 4, 5, 6\}, C = \{1, 2, 4, 5\}$$

We know that $A$ and $B$ are not independent because

$$P(A) = \frac{1}{2}$$
$$P(B) = \frac{5}{6}$$
$$P(A \cap B) = P(\{1, 2\}) = \frac{1}{3}$$

so $P(A) * P(B) = \frac{1}{2} * \frac{5}{6} = \frac{5}{12} =\neq P(A \cap B)$.

We know that $A$ and $B$ are conditionally independent given $C$ because

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(\{1, 2\})}{P(\{1, 2, 4, 5\})} = \frac{1}{2}$$
$$P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{P(\{1, 2, 4, 5\})}{P(\{1, 2, 4, 5\})} = 1$$
$$P((A \cap B)|C) = \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(\{1, 2\})}{P(\{1, 2, 4, 5\})} = \frac{1}{2}$$

so $P((A \cap B)|C) = P(A|C) * P(B|C) = \frac{1}{2}$.

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

**Solution:**

$$A = \{1, 2, 3\}, B = \{1, 2, 4, 5, 6\}, C = \{3, 6\}$$

We showed in part (a) that these same sets $A$ and $B$ were independent.

We know that $A$ and $B$ are not conditionally independent given $C$ because

$$P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(\{3\})}{P(\{3, 6\})} = \frac{1}{2}$$
$$P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{P(\{6\})}{P(\{3, 6\})} = \frac{1}{2}$$
$$P((A \cap B)|C) = \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(\{\})}{P(\{3, 6\})} = 0$$

so $P(A|C) * P(B|C) = \frac{1}{2} * \frac{1}{2} = \frac{1}{4} \neq P((A \cap B)|C)$.

**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?

**Solution:** We will use Bayes' Theorem, which says that

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

Let $B$ be the event that you took the bus, and $A$ be the event that you were late, so that Bayes theorem becomes

$$P(\text{bus}|\text{late}) = \frac{P(\text{late}|\text{bus})P(\text{bus})}{P(\text{late})}.$$

We are given that

$$P(\text{late}|\text{bus}) = 0.5$$
$$P(\text{bus}) = \frac{1}{3}$$

To calculate $P(\text{late})$, notice that since you take exactly one of the three modes of transportation,

$$\begin{aligned} P(\text{late}) &= P(\text{late and bus}) + P(\text{late and drive}) + P(\text{late and walk}) \\ &= P(\text{late}|\text{bus})P(\text{bus}) + P(\text{late}|\text{drive})P(\text{drive}) + P(\text{late}|\text{walk})P(\text{walk}) \\ &= 0.5 * \frac{1}{3} + 0.3 * \frac{1}{3} + 0.05 * \frac{1}{3} \\ &= 0.2833 \end{aligned}$$

Plugging in the values above gives that

$$\begin{aligned} P(\text{bus}|\text{late}) &= \frac{P(\text{late}|\text{bus})P(\text{bus})}{P(\text{late})} \\ &= \frac{0.5 * \frac{1}{3}}{0.2833} \\ &= 0.5882. \end{aligned}$$

Thus, the professor would estimate that there is approximately a 59% chance that you took the bus.

**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?

**Solution:**

We use Bayes Theorem as in part (a), except now

$$\begin{aligned} P(\text{late}|\text{bus}) &= 0.5 \\ P(\text{bus}) &= 0.2 \\ P(\text{late}) &= P(\text{late and bus}) + P(\text{late and drive}) + P(\text{late and walk}) \\ &= P(\text{late}|\text{bus})P(\text{bus}) + P(\text{late}|\text{drive})P(\text{drive}) + P(\text{late}|\text{walk})P(\text{walk}) \\ &= 0.5 * 0.2 + 0.3 * 0.2 + 0.05 * 0.6 \\ &= 0.19 \end{aligned}$$

Plugging in the values above gives that

$$\begin{aligned} P(\text{bus}|\text{late}) &= \frac{P(\text{late}|\text{bus})P(\text{bus})}{P(\text{late})} \\ &= \frac{0.5 * 0.2}{0.19} \\ &= 0.5263. \end{aligned}$$

Thus, the professor would estimate that there is approximately a 53% chance that you took the bus.

**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.

**Solution:**

When performing classification with naive Bayes, we wish to find the label which maximizes the following expression:

$$P(\text{class}|\text{features}) = \frac{P(\text{features}|\text{class}) \cdot P(\text{class})}{P(\text{features})}$$

Note that since $P(\text{features})$ is constant (features are given in the question), we can omit the denominator, and just try to maximize:

$$P(\text{features}|\text{class}) \cdot P(\text{class})$$

The class that maximises this quantity will be our prediction.

In this problem, there are three possible classes, or occupations.

**data scientist (DS):**

$P(\text{DS}) = 10/30$

$P((\text{math, Python, data analysis})|\text{DS}) = 2/10 * 8/10 * 4/10$

Overall, $P((\text{math, Python, data analysis})|\text{DS}) \cdot P(\text{DS}) = 2/10 * 8/10 * 4/10 * 10/30 = \mathbf{0.0213}$

**software engineer (SE):**

$P(\text{SE}) = 11/30$

$P((\text{math, Python, data analysis})|\text{SE}) = 3/11 * 5/11 * 5/11$

Overall, $P((\text{math, Python, data analysis})|\text{SE}) \cdot P(\text{SE}) = 3/11 * 5/11 * 5/11 * 11/30 = \mathbf{0.0207}$

**business analyst (BA):**

$P(\text{BA}) = 9/30$

$P((\text{math, Python, data analysis})|\text{BA}) = 5/9 * 4/9 * 3/9$

Overall, $P((\text{math, Python, data analysis})|\text{BA}) \cdot P(\text{BA}) = 5/9 * 4/9 * 3/9 * 9/30 = \mathbf{0.0247}$

Therefore, the most likely occupation for a mathematics major who likes Python and data analysis is **business analyst**, since it has the largest numerator for the naive Bayes classifier.

**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?

**Solution:**

We compute the numerators only, as in part (a), but this time without incorporating college major.

**data scientist (DS):**

$P(\text{DS}) = 10/30$

$P((\text{Python, data analysis})|\text{DS}) = 8/10 * 4/10$

Overall, $P((\text{Python, data analysis})|\text{DS}) \cdot P(\text{DS}) = 8/10 * 4/10 * 10/30 = \mathbf{0.1067}$

**software engineer (SE):**

$P(\text{SE}) = 11/30$

$P((\text{Python, data analysis})|\text{SE}) = 5/11 * 5/11$

Overall,

$$P((\text{Python, data analysis})|\text{SE}) \cdot P(\text{SE}) = 5/11 * 5/11 * 11/30 = \mathbf{0.0758}$$

**business analyst (BA):**

$P(\text{BA}) = 9/30$

$P((\text{Python, data analysis})|\text{BA}) = 4/9 * 3/9$

Overall,

$$P((\text{Python, data analysis})|\text{BA}) \cdot P(\text{BA}) = 4/9 * 3/9 * 9/30 = \mathbf{0.0444}$$

Therefore, the most likely occupation for a person who likes Python and data analysis is **data scientist** since it has the largest numerator for the naive Bayes classifier. Since college major is a key factor in determining occupation, it makes sense that the classifier makes different predictions when this feature is removed.

**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.

**Solution:**

We would expect that a class that is more represented in the training data would be predicted more often. This happens because the probability of each class is considered in the classifier. For example, if the class distribution in our training data was skewed and there was only one data scientist, we would have $P(\text{data scientist}) = 1/30$ and the numerator of the naive Bayes classifier would be small so we would be unlikely to predict an occupation of data scientist. Note that this may or may not be desirable; for example, if the training data is representative of the underlying overall distribution, then it is reasonable for rare classes to have lower probability. However, if the sample is biased in some way, then it could negatively affect classification accuracy. For example, if our training data is collected through an optional response survey and we happen to not get many responses from data scientists because they are too busy to respond, this will lead to bias against a prediction of data scientist.

**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.

**Solution:**

If there is truly no relationship between this fourth feature and occupation, then there would likely be no large changes to the classifier's outcomes. This extra feature would be "noise", and the classifier would essentially depend on the other 3 features to make predictions. Take the favorite color example. If we are trying to classify a person, and we want to incorporate the fact that their favorite color is red, we'll have an additional term in each of the numerators, which represents the proportion of people in each occupation whose favorite color is purple. Under the assumption that there is no association whatsoever between favorite color and occupation, then the proportion of people in each occupation whose favorite color is purple should be roughly the same across all occupations. So this extra term in each numerator is roughly the same, and doesn't have a substantial effect on which of the numerators is largest. Note that this only works if there really is no association between favorite color and occupation. In practice, many seemingly unrelated features can actually be linked, and this can affect the predictions. For example, it may be the case that favorite color is actually associated with occupation because favorite color is associated with gender (some colors are more popular with one gender) and gender is associated with occupation (some occupations have higher proportions of one gender).

| 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 |