CSCI 5822 Assignment 1

Assigned: Jan 22, 2021. **Due:** Feb 5 (11:59pm, MST), on Canvas.

You have the option to solve the assignment in place and upload the notebook as a PDF (`via "File" -> "Download as" -> "PDF via LaTeX (.pdf)"`) file. OR, use the notebook as a worksheet and upload your answers as a separated, typed PDF file. You may get partial points if you append the notebook to that file.

Do not forget to edit your name here.

Name: Alex Book **Collaborated With:** Cole Sturza

The collaboration/anti-plagiarism policy for this course can be found in the syllabus statements.

The questions provided below will ask you to either write code or write answers in the form of markdown. Markdown syntax guide is here: click here. Using markdown you can typeset formulae using latex. This way you can write nice readable answers with formulae like thus:

> A Markov chain is of the form $p(x_T)\prod_{t=1}^{T-1} p(x_t|x_{t+1})$ for some assignment of the variables to labels x_t.

We will follow the following grading rubrics.

- 10: correct answer, work is shown and clear
- 8: correct answer but work might be difficult to read
- 6: incorrect answer but good attempt
- 4: mediocre attempt or very difficult to read
- 2: problem is written down, no attempt to solve
- 0: problem is not written down

Double click anywhere on this box to find out how your instructor typeset it. Press Shift+Enter to go back.

The Tasks 1-5 are from Excercises 1.1-1.6 of Barber.

Task 1: (10 points). Prove

$$p(x, y|z) = p(x|z)p(y|x, z)$$

and also

$$p(x|y, z) = \frac{p(y|x, z)p(x|z)}{p(y|z)} \, .$$

your answer to task 1

$$\begin{aligned}
p(x, y|z) &= \frac{p(x,y,z)}{p(z)} && \text{Conditional Probability} && (1)\\
&= \frac{p(y,x,z)}{p(z)} && \text{Commutativity of Probability} && (2)\\
&= \frac{p(y|x,z)p(x|z)p(z)}{p(z)} && \text{Chain Rule of Probability} && (3)\\
&= p(y|x,z)p(x|z) && && (4)\\
&= p(x|z)p(y|x,z) && \text{Commutativity of Multiplication} && \square
\end{aligned}$$

$$\begin{aligned}
p(x|y, z) &= \frac{p(x,y,z)}{p(y,z)} && \text{Conditional Probability} && (1)\\
&= \frac{p(x,y|z)p(z)}{p(y,z)} && \text{Chain Rule of Probability} && (2)\\
&= \frac{p(x,y|z)p(z)}{p(y|z)p(z)} && \text{Chain Rule of Probability} && (3)\\
&= \frac{p(x,y|z)}{p(y|z)} && && (4)\\
&= \frac{p(x|z)p(y|x,z)}{p(y|z)} && \text{Proven in first proof of task 1} && (5)\\
&= \frac{p(y|x,z)p(x|z)}{p(y|z)} && \text{Commutativity of Multiplication} && \square
\end{aligned}$$

Task 2: (10 points). Prove the [Bonferroni inequality](https://en.wikipedia.org/wiki/Boole%27s_inequality#Bonferroni_inequalities)

$$p(a, b) \geq p(a) + p(b) - 1 \, .$$

your answer to task 2

Note that $p(a, b) = p(a \cap b)$.

$$\begin{aligned}
p(a \cup b) &= p(a) + p(b) - p(a \cap b) && \text{Addition Rule of Probability} && (1)\\
1 &\geq p(a \cup b) && \text{True by definition, any probability must be less than or equal to 1} && (2)\\
1 &\geq p(a) + p(b) - p(a \cap b) && \text{Combined (1) and (2)} && (3)\\
p(a \cap b) &\geq p(a) + p(b) - 1 && && (4)\\
p(a, b) &\geq p(a) + p(b) - 1 && && \square
\end{aligned}$$

Task 3: (10 points). Consider three variable distributions which admit the factorization

$$p(a, b, c) = p(a|b)p(b|c)p(c) \, ,$$

where all variables are binary. How many parameters are needed to specify distributions of this form?

your answer to task 3

In order to represent the distribution of $p(a|b)$, two parameters are needed. Say, $p(a = 1|b = 0)$ and $p(a = 1|b = 1)$. Using complementation, all other variations of $p(a|b)$ can be found. Similarly, two parameters are needed to represent the distribution of $p(b|c)$. Lastly, only one parameter is needed to represent the distribution of $p(c)$. So, five parameters in total are needed to represent the distributions of the given form.

Task 4: (20 points, 10 each). **(a)** There are two boxes. Box 1 contains three red and five white balls and box 2 contains two red and five white balls. A box is chosen at random $p(b \otimes = 1) = p(b \otimes = 2) = 0.5$ and a ball chosen at random from this box turns out to be red. What is the posterior probability that the red ball came from box 1?
(b) Two balls are placed in a box as follows: A fair coin is tossed and a white ball is placed in the box if a head occurs, otherwise a red ball is placed in the box. The coin is tossed again and a red ball is placed in the box if a tail occurs, otherwise a white ball is placed in the box. Balls are drawn from the box three times in succession (always with replacing the drawn ball back in the box). It is found that on all three occasions a red ball is drawn. What is the probability that both balls in the box are red?

your answer to task 4(a)

Let r represent a ball being red and b represent a ball coming from box 1.

$$p(b|r) = \frac{p(r|b)p(b)}{p(r)} = \frac{\frac{3}{8} * \frac{1}{2}}{(\frac{3}{8} + \frac{2}{7}) * \frac{1}{2}} = \frac{\frac{3}{16}}{\frac{37}{112}} = \frac{21}{37} \approx .568$$

your answer to task 4(b)

Let any capitalized number in text represent that number of red balls drawn in a row (THREE = 3 red balls drawn in a row), RR represent two red balls in the box, WW represent two white balls in the box, and RW represent one ball of each color in the box.

$$\begin{aligned}
p(RR|\text{THREE}) &= \frac{p(\text{THREE}|RR)p(RR)}{p(\text{THREE})}\\
&= \frac{(1)p(RR)}{p(\text{THREE})}\\
&= \frac{(.5)(.5)}{p(\text{THREE}|WW)p(WW) + p(\text{THREE}|RW)p(RW) + p(\text{THREE}|RR)p(RR)}\\
&= \frac{.25}{0 * .25 + .5^3 * .5 + 1 * .25}\\
&= \frac{.25}{.0625 + .25}\\
&= .8
\end{aligned}$$

Task 5: (10 points). A secret government agency has developed a scanner which determines whether a person is a terrorist. The scanner is fairly reliable; 95% of all scanned terrorists are identified as terrorists, and 95% of all upstanding citizens are identified as such. An informant tells the agency that exactly one passenger of 100 aboard an aeroplane in which you are seated is a terrorist. The police haul off the plane the first person for which the scanner tests positive. What is the probability that this person is a terrorist?

your answer to task 5

Let T represent whether or not someone is a terrorist, t represent whether or not someone tested positive for terrorism, and H represent whether or not someone got hauled off the plane.

"Correct" Version

$$\begin{aligned}
p(T = 1|H = 1) &= \Sigma_{i=1}^{100} p(T_i = 1, t_i = 1, t_{1\ldots i-1} = 0|H = 1)\\
&= \Sigma_{i=1}^{100} \frac{p(H = 1|T_i = 1, t_i = 1, t_{1\ldots i-1} = 0)p(T_i = 1, t_i = 1, t_{1\ldots i-1} = 0)}{p(H = 1)}
\end{aligned}$$

$p(H = 1|T_i = 1, t_i = 1, t_{1\ldots i-1} = 0) = 1$, as it is guaranteed that someone is hauled off the plane if the test comes back positive.

$$\begin{aligned}
p(T = 1|H = 1) &= \Sigma_{i=1}^{100} \frac{p(T_i = 1, t_i = 1, t_{1\ldots i-1} = 0)}{p(H = 1)}\\
&= \Sigma_{i=1}^{100} \frac{p(t_i = 1, t_{1\ldots i-1} = 0|T_i = 1)p(T_i = 1)}{p(H = 1)}\\
&= \Sigma_{i=1}^{100} \frac{p(t_i = 1|T_i = 1)p(T_i = 1)\prod_{j=1}^{i-1}(t_j = 0|T_j = 0)}{p(H = 1)}\\
&= \Sigma_{i=1}^{100} \frac{(.95)(.01)\prod_{j=1}^{i-1}(.95)}{p(H = 1)}\\
&= \Sigma_{i=1}^{100} \frac{(.95)(.01)(.95)^{i-1}}{p(H = 1)}\\
&= \frac{(.95)(.01)}{p(H = 1)}\Sigma_{i=1}^{100}(.95)^{i-1}\\
&= \frac{(.95)(.01)}{p(H = 1)}\Sigma_{i=0}^{99}(.95)^{i}\\
&= \frac{(.95)(.01)}{p(H = 1)} \cdot \frac{1 - .95^{100}}{1 - .95}\\
&= \frac{(.95)(.01)(1 - .95^{100})}{(.05)p(H = 1)}\\
&= \frac{(.19)(1 - .95^{100})}{p(H = 1)}\\
&= \frac{(.19)(1 - .95^{100})}{1 - p(H = 0)}\\
&p(H = 0) = .95^{99}(.05), \text{ as it is equal to the probability that all passengers test negative}\\
&= \frac{(.19)(1 - .95^{100})}{1 - (.95^{99}(.05))}\\
&\approx .189
\end{aligned}$$

Breast Cancer Version

$$\begin{aligned}
p(T = 1|t = 1) &= \frac{p(t = 1|T = 1)p(T = 1)}{p(t = 1)} && (1)\\
&= \frac{.95 * .01}{p(t = 1|T = 1) * p(T = 1) + p(t = 1| \sim T = 1) * p(\sim T = 1)} && (2)\\
&= \frac{.95 * .01}{.95 * .01 + .05 * .99} && (3)\\
&= .161 && (4)
\end{aligned}$$

The goal of the following tasks is to give you a bit of practice manipulating data, using Bayes' rule, and constructing a naive Bayes classifier. Naive Bayes is described in `10.1` of Barber and understanding examples `10.1` and `10.2` of the text should help you do this assignment.

Dataset. The [titanic dataset](http://www.cs.toronto.edu/~delve/data/titanic/desc.html) gives the values of four categorical attributes for each of the 2201 people on board the Titanic when it struck an iceberg and sank. The attributes are social class (first class, second class, third class, crew member), age (adult or child), gender, and whether or not the person survived. The titanic dataset is available [here](https://home.cs.colorado.edu/~mozer/Teaching/syllabi/ProbabilisticModels/homework/titanic.txt).

Preparation. Build a joint probability table, like the ones we discussed in class notes, that represents the joint distribution over all variables, i.e., $Pr(\text{Gender}, \text{Age}, \text{Class}, \text{Outcome})$. This table should have 32 entries because $\text{Gender} \in \{\text{male}, \text{female}\}$, $\text{Age} \in \{\text{child}, \text{adult}\}$, $\text{Class} \in \{\text{1st}, \text{2nd}, \text{3rd}, \text{crew}\}$, and $\text{Outcome} \in \{\text{death}, \text{survival}\}$. You will use the data in this table for the following tasks.

Task 6: Probability table (20 points). Build a probability table indicating $Pr(\text{Death} \mid \text{Gender}, \text{Age}, \text{Class})$ for each combination of gender, age, and class. Display this table in the following way:

	Male		Female	
	Child	Adult	Child	Adult
First				
Second				
Third				
Crew				

The rows of each table represent the different classes and the columns the different ages and genders. In each cell of the table, insert the conditional probability. After you've built the probability table, come up with a rule that uses the probabilities to predict death or survival. Then make a second table, a classification table, which lists death or survival for each feature combination. Explain the rule you chose to classify.

Hint. * Since it is [not possible](https://www.tablesgenerator.com/html_tables) to create merged rows with Markdown, this [**Tables Generator**](https://www.tablesgenerator.com/html_tables) may be useful. * Be alert to the possibility of a cell containing no data.

In [2]:
```
# your answer to task 6
import numpy as np
import pandas as pd

titanic_df = pd.read_csv('titanic.txt', delim_whitespace=True, header=None, names = ["Class", "Age", "Gender", "Outcome"])

titanic_df.value_counts(['Class', 'Age', 'Gender','Outcome'])/titanic_df.value_counts(['Class', 'Age', 'Gender'])
```

Out[2]:
```
Class  Age    Gender  Outcome
1st    adult  female  yes        0.972222
                      no         0.027778
              male    no         0.674286
                      yes        0.325714
       child  female  yes        1.000000
              male    yes        1.000000
2nd    adult  female  yes        0.860215
                      no         0.139785
              male    no         0.916667
                      yes        0.083333
       child  female  yes        1.000000
              male    yes        1.000000
3rd    adult  female  no         0.539394
                      yes        0.460606
              male    no         0.837662
                      yes        0.162338
       child  female  no         0.548387
                      yes        0.451613
              male    no         0.729167
                      yes        0.270833
crew   adult  female  yes        0.869565
                      no         0.130435
              male    no         0.777262
                      yes        0.222738
dtype: float64
```

Probability table

	Male		Female	
	Child	Adult	Child	Adult
First	0	.674	0	.028
Second	0	.917	0	.140
Third	.729	.737	.548	.539
Crew	0	.778	0	.130

Classification table

Rule: Anyone with chances of death of greater than a coin flip (greater than .5) is given a 100% chance of death, while anyone with chances of death of less than a coin flip (less than .5) is given a 0% chance of death.

	Male		Female	
	Child	Adult	Child	Adult
First	0	1	0	0
Second	0	1	0	0
Third	1	1	1	1
Crew	0	1	0	0

Task 7: Build a Naive Bayes classifier (20 points). To build the classifier, you must first construct six one-dimensional tables:

- $Pr(\text{Class} \mid \text{death})$
- $Pr(\text{Age} \mid \text{death})$
- $Pr(\text{Gender} \mid \text{death})$
- $Pr(\text{Class} \mid \text{survival})$
- $Pr(\text{Age} \mid \text{survival})$
- $Pr(\text{Gender} \mid \text{survival})$

To be clear on this notation, for $Pr(\text{Age} \mid \text{death})$, your table should have two rows, one for adult and one for child, and you should compute, for each age group, the probability of the deceased being in that age group. Also compute the unconditional probabilities, $Pr(\text{death})$ and $Pr(\text{survival})$, with $Pr(\text{death}) + Pr(\text{survival}) = 1$. From this information, compute $Pr(\text{death} \mid \text{Gender}, \text{Age}, \text{Class})$ using the Naive Bayes assumption. In addition to the probability table, build the classification table as well.

In [58]:
```
# your answer to task 7
num_death = len(titanic_df[(titanic_df['Outcome'] == 'no')])
num_survive = len(titanic_df[(titanic_df['Outcome'] == 'yes')])
pr_death = num_death/(num_death+num_survive)
pr_survive = num_survive/(num_death+num_survive)

classes = titanic_df['Class'].unique()

class_given_death = np.zeros(len(classes))
class_given_survive = np.zeros(len(classes))

for i in range(len(classes)):
    class_given_death[i] = len(titanic_df[(titanic_df['Outcome'] == 'no') & (titanic_df['Class'] == classes[i])])/num_death
    class_given_survive[i] = len(titanic_df[(titanic_df['Outcome'] == 'yes') & (titanic_df['Class'] == classes[i])])/num_surive

d_class = {'Pr(Class | death)': class_given_death, 'Pr(Class | survive)': class_given_survive}
class_df = pd.DataFrame(data=d_class, index=classes)

##########

ages = titanic_df['Age'].unique()

age_given_death = np.zeros(len(ages))
age_given_survive = np.zeros(len(ages))

for i in range(len(ages)):
    age_given_death[i] = len(titanic_df[(titanic_df['Outcome'] == 'no') & (titanic_df['Age'] == ages[i])])/num_death
    age_given_survive[i] = len(titanic_df[(titanic_df['Outcome'] == 'yes') & (titanic_df['Age'] == ages[i])])/num_surive

d_age = {'Pr(Age | death)': age_given_death, 'Pr(Age | survive)': age_given_survive}
age_df = pd.DataFrame(data=d_age, index=ages)

##########

genders = titanic_df['Gender'].unique()

gender_given_death = np.zeros(len(genders))
gender_given_survive = np.zeros(len(genders))

for i in range(len(genders)):
    gender_given_death[i] = len(titanic_df[(titanic_df['Outcome'] == 'no') & (titanic_df['Gender'] == genders[i])])/num_death
    gender_given_survive[i] = len(titanic_df[(titanic_df['Outcome'] == 'yes') & (titanic_df['Gender'] == genders[i])])/num_surive

d_gender = {'Pr(Gender | death)': gender_given_death, 'Pr(Gender | survive)': gender_given_survive}
gender_df = pd.DataFrame(data=d_gender, index=genders)

##########

death_given_class_age_gender = np.zeros((len(classes), len(ages), len(genders)))

for i in range(len(classes)):
    for j in range(len(ages)):
        for k in range(len(genders)):
            numerator = class_given_death[i]*age_given_death[j]*gender_given_death[k]*pr_death
            denominator = numerator + class_given_survive[i]*age_given_survive[j]*gender_given_survive[k]*pr_survive

            death_given_class_age_gender[i, j, k] = numerator/denominator

# classes, ages, genders
# death_given_class_age_gender

first_adult_male = death_given_class_age_gender[0,0,0]
first_adult_female = death_given_class_age_gender[0,0,1]
first_child_male = death_given_class_age_gender[0,1,0]
first_child_female = death_given_class_age_gender[0,1,1]
second_adult_male = death_given_class_age_gender[1,0,0]
second_adult_female = death_given_class_age_gender[1,0,1]
second_child_male = death_given_class_age_gender[1,1,0]
second_child_female = death_given_class_age_gender[1,1,1]
third_adult_male = death_given_class_age_gender[2,0,0]
third_adult_female = death_given_class_age_gender[2,0,1]
third_child_male = death_given_class_age_gender[2,1,0]
third_child_female = death_given_class_age_gender[2,1,1]
crew_adult_male = death_given_class_age_gender[3,0,0]
crew_adult_female = death_given_class_age_gender[3,0,1]
crew_child_male = death_given_class_age_gender[3,1,0]
crew_child_female = death_given_class_age_gender[3,1,1]

# print('first_adult_male:', first_adult_male)
# print('first_adult_female:', first_adult_female)
# print('first_child_male:', first_child_male)
# print('first_child_female:', first_child_female)
# print('second_adult_male:', second_adult_male)
# print('second_adult_female:', second_adult_female)
# print('second_child_male:', second_child_male)
# print('second_child_female:', second_child_female)
# print('third_adult_male:', third_adult_male)
# print('third_adult_female:', third_adult_female)
# print('third_child_male:', third_child_male)
# print('third_child_female:', third_child_female)
# print('crew_adult_male:', crew_adult_male)
# print('crew_adult_female:', crew_adult_female)
# print('crew_child_male:', crew_child_male)
# print('crew_child_female:', crew_child_female)
```

Probability table

	Male		Female	
	Child	Adult	Child	Adult
First	.317	.528	.044	.099
Second	.522	.725	.097	.206
Third	.696	.847	.184	.352
Crew	.710	.855	.195	.368

Classification table

Rule: Anyone with chances of death of greater than a coin flip (greater than .5) is given a 100% chance of death, while anyone with chances of death of less than a coin flip (less than .5) is given a 0% chance of death.

	Male		Female	
	Child	Adult	Child	Adult
First	0	1	0	0
Second	1	1	0	0
Third	1	1	0	0
Crew	1	1	0	0

Task 8: Discussion (10 points, bonus) The classification tables you built in Tasks 6 and 7 are not identical. Discuss the advantages/disadvantages of each table for making predictions in case of another disaster like the Titanic (assuming it occurred at the same time in history). Under what circumstances would you expect an empirical table to provide better predictions? Under what circumstances would you expect the naive Bayes table to provide better predictions?

your answer to task 8

An empirical table may provide better predictions when there are similar passenger distributions as there were on Titanic (which seems to have been a fairly typical distribution). For example, it works better when there are no child crew members (which one would hope to be true in all cases).

A naive Bayes table may be of more use when there are more even (and likely unorthodox) distributions of passengers and crew. For example, it would work better if the probability of a crew member being a child were equal to that of a crew member being an adult. In such a case, the Naive Bayes assumption that all given attributes have independent probabilities would be correct, and thus wouldn't harm the accuracy of its predictions.