IS5152 Data-driven Decision Making
SEMESTER II 2022-2023
Assignment 3
Due: Friday, 24 March 2023, 11.59pm

Instructions:

- Upload your answer as a single pdf file to Canvas.
- Name your file according to your student ID number, e.g. Asg3-A1234567X.pdf.

1. (10 points) A group of dentists is considering opening of a new private clinic. If the demand for dentists is high (that is, there is a favorable market for the clinic), the dentists could realize a net profit of $1,000,000. If the market is not favorable, they would lose $400,000. If they do not proceed at all, there will be no cost/profit. In the absence of any market data, the best the dentists can guess is that there is a 50-50 chance the clinic will be successful. The dentists may engage a market research firm to perform a study of the market, at a fee of $50,000. The market researchers claim their past experience shows that when the markets were favorable, their study correctly predicted success 70% of the time. Thirty percent of the time the study falsely predicted a failure. On the other hand, when the market condition was unfavorable, the study was correct 80% of the time in predicting a failure. The remaining 20% of the time, it incorrectly predicted a success.

   (a) (2 points) State the Expected Value of Perfect Information.

   (b) (6 points) State the Expected Value of Sample Information.

   (c) (2 points) What is the best decision that the dentists should take?

2. (10 points) The following table consists of training data from an employee dataset. The attribute *department* has 3 possible values: sales, systems, marketing. The attribute *age* has 4 possible values: [21 to 25], [26 to 30], [31 to 35], [36 to 40]. The attribute *salary* has 3 possible values: $< 30K$, $[30K$ to $40K]$, $> 40K$. The target attribute *status* is binary-valued: junior or senior.

department	age	salary	status
sales	[31 to 35]	$> 40K$	senior
sales	[26 to 30]	$[30K - 40K]$	junior
systems	[21 to 25]	$< 30K$	junior
systems	[36 to 40]	$> 40K$	senior
systems	[26 to 30]	$> 40K$	junior
systems	[31 to 35]	$[30K - 40K]$	junior
marketing	[31 to 35]	$[30K - 40K]$	senior
marketing	[36 to 40]	$[30K - 40K]$	senior
marketing	[26 to 30]	$> 40K$	senior
marketing	[21 to 25]	$[30K - 40K]$	junior

(a) (4 points) What would be the naive Bayes prediction for the status of an employee with the attribute values: (marketing,[31 to 35], $> 40K$)?

(b) (6 points) Laplace smoothing is applied to the data with parameter k set to 3. What is the predicted status for an employee with attribute values: (sales,[36 to 40],$< 30K$)?

3. (10 points) Consider the data set below which consists of records from the previous 20 days when we played/did not play tennis.

Day	Outlook	Temperature	Humidity	WindCondition	PlayTennis
D1	Sunny	32	90	Strong	NO
D2	Sunny	34	90	Strong	NO
D3	Overcast	19	80	Weak	YES
D4	Rain	25	86	Strong	YES
D5	Rain	22	85	Weak	YES
D6	Sunny	28	75	Strong	NO
D7	Overcast	21	80	Weak	YES
D8	Sunny	29	67	Strong	NO
D9	Sunny	30	78	Weak	YES
D10	Rain	27	80	Weak	NO
D11	Rain	25	82	Weak	YES
D12	Overcast	22	86	Strong	YES
D13	Overcast	20	76	Weak	YES
D14	Rain	33	90	Strong	NO
D15	Sunny	23	90	Strong	NO
D16	Rain	28	90	Strong	NO
D17	Overcast	20	76	Weak	YES
D18	Overcast	22	76	Weak	YES
D19	Sunny	21	86	Weak	YES
D20	Overcast	29	75	Weak	YES

The categorical attribute Outlook has 3 possible values: Sunny, Overcast or Rain. The

categorical attribute WindCondition has 2 possible values: Weak or Strong. Temperature (in Celsius) and Humidity (in percent) are continuous input attributes.

The final logistic regression model after feature selection has just 2 input variables:

```
Coefficients:
                    Estimate     Std. Error     z value   Pr(>|z|)
    (Intercept)      10.7385         6.7215       1.598   0.1101
    Temperature      -0.4530         0.2641      -1.715   0.0863 .
    WindWeak          3.2365         1.6869       1.919   0.0550 .
    ---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Null deviance: xxxxx  on 19  degrees of freedom
Residual deviance: 11.211  on 17  degrees of freedom
AIC: xxxxx
```

(a) (2 points) Compute and interpret the odds ratio point estimates for the variables Temperature and WindWeak.

Note: The variable WindWeak = 1 if the WindCondition is Weak, WindWeak = 0 if WindCondition = Strong.

(b) (2 points) Compute the values of the Null deviance and AIC.

(c) (2 points) The predicted output for the 20 data samples are as follows:

$D1 - D7:$ 0.0228, 0.0093, 0.9954, 0.3572, 0.9822, 0.1249, 0.9886,

$D8 - D14:$ 0.0832, 0.5948, 0.8511, 0.9340, 0.6839, 0.9927, 0.0146,

$D15 - D20:$ 0.5790, 0.1249, 0.9927, 0.9822, 0.9886, 0.6976

Compute the percent concordant and discordant.

(d) (2 points) What is the highest classification accuracy that can be achieved by this model?

(e) (2 points) Day 21 is predicted to be Sunny, with Weak Wind condition, 25 degrees C and 80% humidity. Are we playing tennis?