IS5152 Data-driven Decision Making
SEMESTER II 2022-2023
Assignment 3 - Suggested solution

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.

   - EV with PI = 0.5(1,000,000) + 0.5(0) = 500,000
   - EV with OI = 0.5(1,000,000) + 0.5(-400000) = 300,000
   - EV of PI = EVwPI - EVwOI = 200,000.

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

   Define: FM = favourable market, UM = unfavorable market, FS = favorable study, US = unfavorable study

   Conditional probabilities:

$$\begin{aligned} P(FS|FM) &= 0.70, \qquad & P(US|UM) &= 0.8 \\ P(US|FM) &= 0.30, \qquad & P(FS|UM) &= 0.2 \end{aligned}$$

   Original probabilities: $P(FM) = 0.5, P(UM) = 0.5$

   Joint probabilities:

$$\begin{aligned} P(FS \text{ and } FM) = P(FS|FM) \times P(FM) &= 0.35 \\ P(FS \text{ and } UM) = P(FS|UM) \times P(UM) &= 0.10 \\ P(US \text{ and } FM) = P(US|FM) \times P(FM) &= 0.15 \\ P(US \text{ and } UM) = P(US|UM) \times P(UM) &= 0.40 \end{aligned}$$

   Marginal probabilities: P(US) = 0.55, P(FS) = 0.45

Revised probabilities:

$$\begin{aligned}
P(FM|FS) = P(FM \text{ and } FS)/P(FS) &= 35/45 = 7/9 \\
P(UM|FS) = P(UM \text{ and } FS)/P(FS) &= 10/45 = 2/9 \\
P(FM|US) = P(FM \text{ and } US)/P(US) &= 15/55 = 3/11 \\
P(UM|US) = P(UM \text{ and } US)/P(US) &= 40/55 = 8/11
\end{aligned}$$

With sample information:

If the study indicates favorable market:
expected return $= (7/9) \times 1,000,000 + (2/9) \times (-400,000) = 688888.89$

If the study indicates unfavorable market:
expected return $= (3/11) \times 1,000,000 + (8/11) \times (-400,000) = -18,181.8$

Expected value with sample information $= 0.45 \times 688,888.89 + 0.55 \times 0 = 310,000$

Expected value with original information $= 0.5 \times 1,000,000 + 0.5 \times (-400,000) = 300,000$

Expected value of sample information = 310,000 - 300,000 = 10,000

(c) (2 points) What is the best decision that the dentists should take? Since the expected value of sample information is less than the fee for the study, the best decision is to open the clinic without conducting a study of the market condition.

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], $> 40K$)?

- P(senior) = 5/10, P(junior) = 5/10
- P(marketing|senior) = 3/5
- P([31,35]|senior) = 2/5

- P($> 40K$|senior) = 3/5
- P(senior)×P(marketing|senior)×P([31,35]|senior)×P($> 40K$|senior) = 9/125 = 0.072
- P(junior) = 5/10, P(junior) = 5/10
- P(marketing|junior) = 1/5
- P([31,35]|junior) = 1/5
- P($> 40K$|junior) = 1/5
- P(junior)×P(marketing|junior)×P([31,35]|junior)×P($> 40K$|junior) = 1/250 = 0.004
- Predict: senior

(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],$< 30K$)?

- original: P(sales|senior) = 1/5
- original: P([36-40]|senior) = 2/5
- original: P($< 30K$|senior) = 0
- original: P(sales|junior) = 1/5
- original: P([36-40]|junior) = 0
- original: P($< 30K$|junior) = 1/5
- department: 3 levels, P(sales|senior) = 4/(5 + 9) = 4/14
- department: 3 levels, P(sales|junior) = 4/(5 + 9) = 4/14
- age: 4 levels, P([36-40]|senior) = 5/(5 + 12) = 5/17
- age: 4 levels, P([36-40]|junior) = 3/(5 + 12) = 3/17
- salary: 3 levels, P($< 30K$|senior) = 3/(5 + 9) = 3/14
- salary: 3 levels, P($< 30K$|junior) = 4/(5 + 9) = 4/14
- For senior: P(senior)×P(sales|senior)×P([36-40]|senior)×P($< 30K$|senior) = (5/10) × (4/14) × (5/17) × (3/14) = 15/1666 = 0.00900
- For junior: P(junior)×P(sales|junior)×P([36-40]|junior)×P($< 30K$|junior) = (5/10) × (4/14) × (3/17) × (4/14) = 12/1666 = 0.00720
- Predict senior

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.

- Temperature: $\beta = -0.4530, e^{\beta} = 0.6357$: A one degree C increase in Temperature, decreases the odds of PlayTennis = Yes by about 36.5% (Wind condition remains the same).

- WindWeak: $\beta = 3.2365, e^{\beta} = 25.44$: For 2 days with the same temperature, the day when WindCondition is weak, the odds of PlayTennis = Yes is more than 25 times the odds of the day with strong (= not weak) WindCondition.

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

- Null deviance $= -2\,(12\log_e(0.6) + 8\log_e(0.4)) = 26.920$.
- AIC = Residual deviance $+2 \times (2+1) = 11.211 + 6 = 17.211$

(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.

- There are 12 YES and 8 NO samples, the number of pairs = 12*8 = 96
- Sort the predicted output:

Day	PlayTennis	pred
2	NO	0.0093
14	NO	0.0146
1	NO	0.0228
8	NO	0.0832
6	NO	0.1249
16	NO	0.1249
4	YES	0.3572
15	NO	0.5790
9	YES	0.5948
12	YES	0.6839
20	YES	0.6978
10	NO	0.8511
11	YES	0.9340
5	YES	0.9822
18	YES	0.9822
7	YES	0.9886
19	YES	0.9886
13	YES	0.9927
17	YES	0.9927
3	YES	0.9954

- Pairs (4,15), (4,10), (9,10), (12,10), (20,10) are discordant.
- Discordance = (5/96)(100%) = 5.21%.
- Concordance = 100% - 5.21% = 94.79%

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

- Threshold = (0.1249 + 0.3572)/2: # misclassifications = 2 NO
- Threshold = (0.3572 + 0.5790)/2: # misclassifications = 1 YES + 2 NO
- Threshold = (0.5790 + 0.5948)/2: # misclassifications = 1 YES + 1 NO
- Threshold = (0.6978 + 0.8511)/2: # misclassifications = 4 YES + 1 NO
- Threshold = (0.8511 + 0.9340)/2: # misclassifications = 4 YES

Answer: Best accuracy = (20-2)/20 = 0.9

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

- Predicted output: $1/(1+e^{-(w_0+w_1x_1+w_2x_2)}) = 1/(1+e^{-(10.738-0.453(25)+3.237(1))}) = 1/(1+e^{-2.65}) = 0.9340$.
- Predict: YES.