IS5152 Assignment 3 Wang Ruicong A0244297W

1. (a) EVWPI $= 0.5 \times [\max(1{,}000{,}000, 0)] + 0.5 \times [\max(-400000, 0)]$

$= 0.5 \times 1000000 + 0.5 \times 0 = 500000$

EVWOI $= \max(0.5 \times 1000000 - 0.5 \times 400000, 0) = \max(300000, 0)$

$= 300000$

Therefore, the expected value of perfect information is

EVPI $=$ EVWPI $-$ EVWOI $= 500000 - 300000 = 200000$

(b) We know that $P(LS|NS) = 0.7$ $\quad P(LF|NF) = 0.8$ $\quad P(NS) = 0.5$

$P(LF|NS) = 0.3$ $\quad P(LS|NF) = 0.2$ $\quad P(NF) = 0.5$

$\therefore P(NS \cap LS) = 0.5 \times 0.7 = 0.35$ $\quad P(NS \cap LF) = 0.5 \times 0.3 = 0.15$

$P(NF \cap LS) = 0.5 \times 0.2 = 0.1$ $\quad P(NF \cap LF) = 0.5 \times 0.8 = 0.4$

$\therefore P(LS) = 0.35 + 0.1 = 0.45$ $\quad P(LF) = 0.15 + 0.4 = 0.55$

$\therefore P(NS|LS) = \frac{0.35}{0.45} = \frac{7}{9}$ $\quad P(NF|LS) = \frac{0.1}{0.45} = \frac{2}{9}$

$P(NS|LF) = \frac{0.15}{0.55} = \frac{3}{11}$ $\quad P(NF|LF) = \frac{0.4}{0.55} = \frac{8}{11}$

Local success and market nationally

Expected return: $\frac{7}{9} \times 1000000 - \frac{2}{9} \times 400000 - 50000 = \frac{5750000}{9}$

Local success and does not market nationally

Expected return: -50000

Local Failure and market nationally

Expected return: $\frac{3}{11} \times 1000000 - \frac{8}{11} \times 400000 - 50000 = -\frac{750000}{11}$

Local Failure and do not market nationally

Expected return: $-50000 > -\frac{750000}{11}$

Therefore, the overall expected return with research is

$\frac{9}{20} \times \frac{5750000}{9} - \frac{11}{20} \times 50000 = 287500 - 27500 = 260000$

If the research is costless, the expected return is $260000 + 50000 = 310000$

$\therefore$ EVWSI $= \max(310000, 300000) = 310000$

Therefore, the Expected Value of sample information is

EVSI $=$ EVWSI $-$ EVWOI $= 310000 - 300000 = 10000$

(c) The EVSI is less than the cost of market research $(10000 < 50000)$

Therefore, the dentists should not engage a market research and directly open a new private clinic

2. (a) $P(\text{senior}) = \frac{5}{10} = \frac{1}{2}$ $\quad P(\text{junior}) = \frac{5}{10} = \frac{1}{2}$

$P(\text{marketing} \mid \text{senior}) = \frac{3}{5}$ $\quad P(\text{marketing} \mid \text{junior}) = \frac{1}{5}$

$P([31 \text{ to } 35] \mid \text{senior}) = \frac{2}{5}$ $\quad P([31 \text{ to } 35] \mid \text{junior}) = \frac{1}{5}$

$P(>40K \mid \text{senior}) = \frac{3}{5}$ $\quad P(>40K \mid \text{junior}) = \frac{1}{5}$

For senior $\frac{1}{2} \times \frac{3}{5} \times \frac{2}{5} \times \frac{3}{5} = \frac{9}{125}$

For junior $\frac{1}{2} \times \frac{1}{5} \times \frac{1}{5} \times \frac{1}{5} = \frac{1}{250}$

$\frac{9}{125} > \frac{1}{250}$ $\Rightarrow$ The naive Bayes prediction is senior

(b) After Laplace smoothing with $K=3$

$P(\text{sales} \mid \text{senior}) = \frac{1+3}{5+3\times3} = \frac{2}{7}$ $\quad P(\text{sales} \mid \text{junior}) = \frac{1+3}{5+3\times3} = \frac{2}{7}$

$P([36 \text{ to } 40] \mid \text{senior}) = \frac{2+3}{5+3\times4} = \frac{5}{17}$ $\quad P([36 \text{ to } 40] \mid \text{junior}) = \frac{0+3}{5+3\times4} = \frac{3}{17}$

$P(<30K \mid \text{senior}) = \frac{0+3}{5+3\times3} = \frac{3}{14}$ $\quad P(<30K \mid \text{junior}) = \frac{1+3}{5+3\times3} = \frac{2}{7}$

For senior $\frac{1}{2} \times \frac{2}{7} \times \frac{5}{17} \times \frac{3}{14} = \frac{15}{1666}$

For junior $\frac{1}{2} \times \frac{2}{7} \times \frac{3}{17} \times \frac{2}{7} = \frac{12}{1666} = \frac{6}{833}$

$\frac{15}{1666} > \frac{6}{833}$ $\Rightarrow$ The naive Bayes prediction status is senior

3. (a) The odds ratio for temperature is $e^{-0.4530} \approx 0.6357$

It means when the temperature increase 1 (celsius)

The odds of playing tennis decrease to 63.57% of original odds

The odds ratio of WindWeak is $e^{3.2365} \approx 25.4445$

~~It means that when the wind condition changes from weak to strong~~
~~The odds of playing tennis increase to 25.4445 times of original odds~~

It means that the odds of playing tennis when the wind condition is week is more than 25-fold higher than when the wind condition is strong

(b) $\hat{\pi} = \frac{12}{20} = 0.6$ $\quad \ln g(Y_1, Y_2, \cdots Y_n) = 12\ln(\frac{12}{20}) + 8\ln(\frac{8}{20}) \approx -13.46$

Null deviance $= (-2) \times (-13.46) = 26.92$

AIC $= 11.211 + 2 \times (2+1) = 17.211$

(c) We sort the predicted values first in ascending order

	D2	D14	D1	D8	D6	D16	D4
predict	0.0093	0.0146	0.0228	0.0832	0.1249	0.1249	0.3572
true	No	No	No	No	No	No	Yes

	D15	D9	D12	D20	D10	D11
predict	0.5790	0.5948	0.6839	0.6976	0.8511	0.9340
true	No	Yes	Yes	Yes	No	Yes

	D5	D18	D7	D19	D13	D17	D3
predict	0.9822	0.9822	0.9886	0.9886	0.9927	0.9927	0.9954
true	Yes	Yes	Yes	Yes	Yes	Yes	Yes

There are total $12 \times 8 = 96$ pairs while the number of discordant pairs is $1+4=5$

The percent concordant is $\frac{96-5}{96} = \frac{91}{96} \approx 0.948$

The percent discordant is $\frac{5}{96} \approx 0.052$

(d) The highest classification accuracy is 90%

If we set a threshold to 0.35 or 0.58

we can successfully predict 18 out of 20

Therefore, the highest classification accuracy is $\frac{18}{20} = 90\%$

(e) $$P(\text{Yes}) = \frac{1}{1+e^{-(10.7385-0.4530\times 25+3.2365)}}$$

$$= \frac{1}{1+e^{-2.65}} \approx 0.934 > 0.5$$

Therefore, we will choose to play tennis