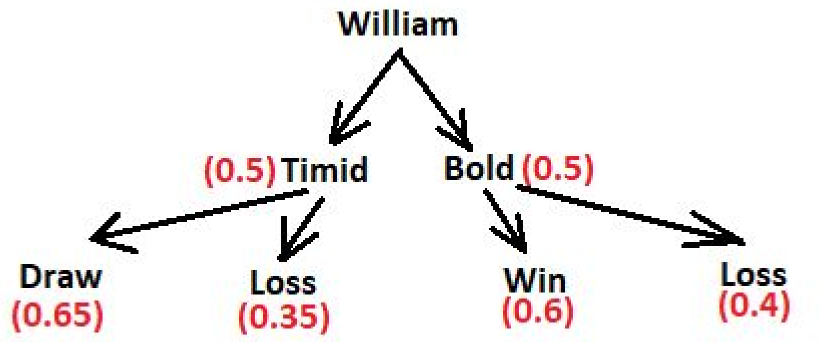
CS 464 MACHINE LEARNING HW1

Name: Fatih Sevban Uyanık
Id: 21602486
Section: 2

Question 1 → The Chess Game [15 pts]



During Sudden Death ⇒ William plays 100% bold.
Before Sudden Death ⇒ William plays 50% bold and 50% timid.

Events
W: William wins the match.
W_i: Willam wins the i'th game.
L_i: William loses the i'th game.
D_i: William draws the i'th game.
B: William plays bold in all matches.
B_i: William plays bold in the i'th game.
T: William plays Timid in the first two games.
T_i: William plays timid in the i'th game.
S: Steve guesses correctly.

a)

$$\begin{aligned}\mathbf{P(W|B)} &= P(L1 \cap W2 \cap W3 \mid B) + P(W1 \cap L2 \cap W3) + P(W1 \cap W2 \mid B) \\ &= P(L1 \mid B1) \,.\, P(W2 \mid B2) \,.\, P(W3 \mid B3) + P(W1 \mid B1) \,.\, P(L2 \mid B2) \,.\, P(W3 \mid B3) \\ &\quad + P(W1 \mid B1) \,.\, P(W2 \mid B2) \\ &= (0.4) \,.\, (0.6) \,.\, (0.6) + (0.6) \,.\, (0.4) \,.\, (0.6) + (0.6) \,.\, (0.6) \\ &= \mathbf{0.648}\end{aligned}$$

b)

P(W|T) = 0

By playing timid, William can't win because by playing constantly timid, he has only a chance of drawing and losing. Thats why, the probability of winning by playing timid in all first three games is 0.

c)

P(B1 | L1) = $(P(L_1|B_1) \,.\, P(B_1)) / P(L_1)$

$= (P(L_1|B_1) \,.\, P(B_1)) / (P(T \cap L_1) + P(B \cap L_1))$

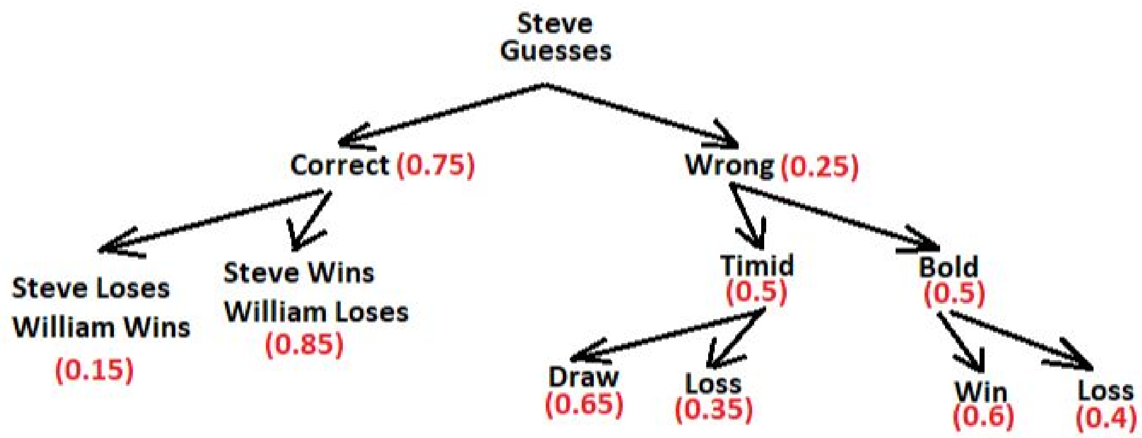
$= ((0.4)(0.5)) / ((0.5)(0.4) + (0.5)(0.35))$

$= (0.2) / (0.2 + 0.175)$

= 0.533333

d)

Guess 1 → William plays Timid. → Steve plays Bold.
Guess 2 → William plays Bold. → Steve plays Timid.



P(L$_i$) = $P(L_i \cap S) + P(L_i \cap S')$

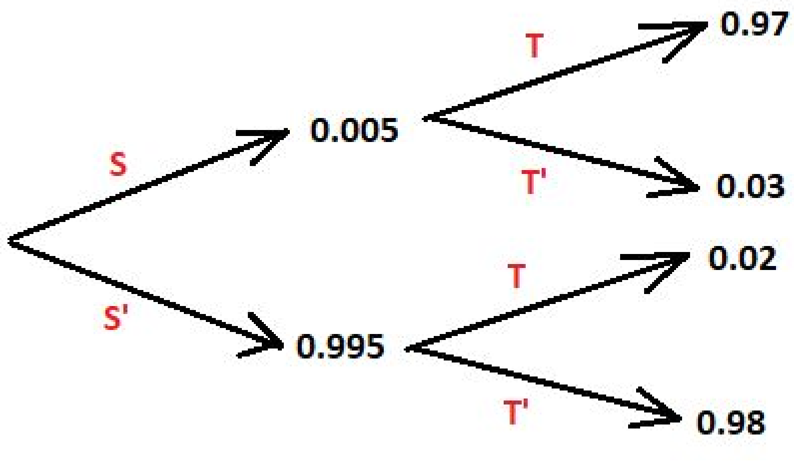
$= P(S) \,.\, P(L_i \mid S) + P(S') \,.\, P(L_i \mid S')$

$= (0.75)\,(0.85) + (0.25) \,.\, [\, (0.5)(0.35) + (0.5)(0.4) \,]$

= 0.73125

Question 2 → Medical Diagnosis [10 pts]

S : A person that is chosen from the population has the disease.
T : Test applied is positive that reflects the person has the disease.



2.1
P(S) = 0.005
P(S') = 0.995
P(T|S) = 0.97
P(T'|S) = 0.03
P(T'|S') = 0.98
P(T|S') = 0.02

2.2
It is given that the Test is positive. In this case, we need to find the probabilities **P(S|T)** and **P(S'|T)** because the test may be positive in case of the patient has the disease or not has the disease.

P(S|T) = (P(T|S) . P(S)) / P(T)
= (P(T|S) . P(S)) / (P(T∩S) + P(T∩S'))
= (P(T|S) . P(S)) / (P(S) . P(T|S) + P(S') . P(T|S'))
= ((0.97) . (0.05)) / ((0.05) . (0.97) + (0.995) . (0.02))
= 0.71

P(S'|T) = (P(T|S') . P(S')) / P(T)
= (P(T|S') . P(S')) / (P(T∩S) + P(T∩S'))
= (P(T|S') . P(S')) / (P(S) . P(T|S) + P(S') . P(T|S'))
= ((0.02) . (0.995)) / ((0.05) . (0.97) + (0.995) . (0.02))
= 0.29

Because P(S/T) is higher, the patient is more likely to have the disease. That's why, the patient should be diagnosed as having the disease.

Question 3 → MLE and MAP [15 pts]

$$\mathbf{P}(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

3.1

P(D|λ) = P(x_1, x_2, x_3, ….. , x_n | D)
= P(x_1|λ) . P(x_2|λ) . P(x_3|λ) ……...P(x_n|λ)
= ((e^-λ . λ^x_1) / x_1!) . ((e^-λ . λ^x_2) / x_2!) . ((e^-λ . λ^x_3) / x_3!) ……...((e^-λ . λ^x_n) / x_n!)
= (e^-λ.n)(λ^x_1.x_2.x_3….x_n) / (x_1!x_2!x_3!.x_n!)

λ̂ = arg max ln(P(D|λ))
= **arg max** ln((e^-λ.n)(λ^x_1.x_2.x_3….x_n) / (x_1!x_2!x_3!.x_n!))
= **arg max** (-λn).ln(e) + (x_1+x_2+x_3+....+x_n).ln(λ) - ln(x_1!x_2!x_3!.x_n!)

⇒ taking the derivative of ln(P(D|λ))
⇒ d/dλ [ln(P(D|λ))] = d/dλ [-λn + (x_1+x_2+x_3+....+x_n).ln(λ) - ln(x_1!x_2!x_3!.x_n!)]
= -n + ((x_1+x_2+x_3+....+x_n) / λ)

⇒ equating the derivative to 0, we get

-n + ((x_1+x_2+x_3+....+x_n) / λ) = 0
(x_1+x_2+x_3+....+x_n) / λ = n
λ = (x_1+x_2+x_3+....+x_n) / n is the MLE of λ

3.2

$$Gamma(x \mid \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$$

λ̂MAP = arg max ln(P(λ|D))
= arg max ln(P(D|λ) . P(λ) / P(D))

$$= \textbf{arg max} \quad \ln\left(\frac{\prod_{i=1}^{n} (\frac{e^{-\lambda}.\lambda^{x_i}}{x_i!}).(\frac{\beta^\alpha.\lambda^{\alpha-1}.e^{-\lambda\beta}}{\Gamma(\alpha)})}{(\prod_{i=1}^{n} \frac{\beta^\alpha x_i^{\alpha-1} e^{-\beta x_i}}{\Gamma(\alpha)})} \right)$$

$$= \textbf{arg max} \quad \ln(\prod_{i=1}^{n} \frac{e^{-\lambda}.\lambda^{x_i}}{x_i!}) + \ln(\frac{\beta^\alpha \lambda^{\alpha-1} e^{-\lambda\beta}}{(\alpha-1)!}) - \ln(\prod_{i=1}^{n} \frac{\beta^\alpha x_i^{\alpha-1} e^{\beta x_i}}{(\alpha-1)!})$$

= arg max $\lambda n + (\sum_{i=1}^{n} x_i).\ln(\lambda) - \sum_{i=1}^{n}\ln(x_i!) + \alpha\ln(\beta) + (\alpha-1)\ln\lambda - \lambda\beta - \ln((\alpha-1)!)$

$$-\sum_{i=1}^{n}\ln(\frac{\beta^{\alpha} x_i^{\alpha-1} e^{\beta x_i}}{(\alpha-1)!})$$

⇒ taking the derivative of ln(P(λ|D))

$$\frac{d}{d\lambda}(P(\lambda \mid D)) = n + \frac{\sum_{i=1}^{n} x_i}{\lambda} + \frac{(\alpha-1)}{\lambda} - \beta$$

⇒ equating derivative to 0

$$n + \frac{\sum_{i=1}^{n} x_i}{\lambda} + \frac{(\alpha-1)}{\lambda} - \beta = 0$$

$$\lambda = \frac{\sum_{i=1}^{n} x_i + (\alpha-1)}{\beta - n}$$

3.3

$$\lambda \sim U(a,b) \Rightarrow P(\lambda \mid a,b) = \frac{1}{b-a}$$

$$\Rightarrow P(\lambda \mid D) = \frac{P(D \mid \lambda)P(\lambda)}{P(D)}$$

$$= \frac{P(x_1 \mid \lambda).P(x_2 \mid \lambda).......P(x_n \mid \lambda)P(\lambda)}{P(x_1,x_2,x_3.....x_n)}$$

$$= \frac{\left[\prod_{i=1}^{n}\frac{\lambda^{x_i}.e^{-\lambda}}{x!}\right].(\frac{1}{b-a})}{n.(\frac{1}{b-a})}$$

$$= \frac{\prod_{i=1}^{n}\frac{\lambda^{x_i}.e^{-\lambda}}{x!}}{n}$$

$$\lambda_{MAP} = \arg\max P(\lambda \mid D)$$
$$= \arg\max \ln\left[\frac{\prod_{i=!}^{n}(\frac{\lambda^{x_i}.e^{-\lambda}}{x!})}{n}\right]$$
$$= \arg\max \sum_{i=1}^{n} x_i.\ln(\lambda) - \sum_{i=1}^{n}(\lambda) - \sum_{i=1}^{n}(x!) - \ln(n)$$

⇒ taking the derivative of ln(P(λ|D))

$$\frac{d}{d\lambda}(\ln P(\lambda \mid D)) = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n$$

⇒ equating derivative to 0

$$\frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0$$
$$\lambda_{MAP} = \frac{\sum_{i=1}^{n} x_i}{n}$$

As it can be seen according to the result and 3.1, MLE estimate of λ and MAP estimate of λ with uniform prior U(a; b) is the same for any a and b.

Question 4 → Sentiment Analysis on Tweets [60 pts]

4.1

Because in naive bayes, we are interested in classifying the tweets. In each tweet class (positive, negative, neutral), we have the same denominator and the denominator is a constant number with respect to λ. That is to say, the class that has the biggest MLE won't change. Only the value of MLE might change, however, the class that have the maximum MLE would be the same. Because our purpose is classifying and finding the class that has the maximum MIE, the denominator can be ignored.

4.2

Negative Tweet Count in Training Set : 7091
Positive Tweet Count in Training Set : 2004
Neutral Tweet Count in Training Set : 2617
SUM : 11712

Percentage Of Negative Tweet Count = ((7091).(100)) / (11712) = %60.054474044
Percentage Of Positive Tweet Count = ((2004).(100)) / (11712) = %17.11065574
Percentage Of Neutral Tweet Count = ((2617).(100)) / (11712) = %22.34460383

- Percentage Of Negative Tweet Count is **%60.054474044**
- Our training set is **skewed** towards the **negative class.**
- Having an imbalanced training dataset can affect my model because the model may perform well on classifying negative tweets, however, it may classify wrong the neutral and positive tweets because our model is not trained very well for the negative and neutral tweets.
- The solution for having an unbalanced tree can be **oversampling** and **undersampling**. In over-sampling, positive and neutral tweets for the training set can be increased. On the other hand, in under sampling, the number of negative tweets can be reduced in the training data set. İf possible, getting rid of redundant tweets in negative tweets can even improve our model further.

4.3

Since there are V words and 3 classes, **3V-1** parameters needed to be estimated. Since V consists of 5722 words, 3.(5722) - 1 = **17165** words needed to be estimated.

Notice: The remaining questions are in the notebooks. The comments and answers are given there also.