COM6101 Assignment 2

P233340 Kwok Tsz Yi

Part 1

1. CGPA has a 0.1194 positive correlation with the chance of admission
2. Predictive chance:
   0.0019 * 327 + 0.0029 * 115 + 3*0 + 0.0179 * 3.5 + 8.2 * 0.1194 + 0 – 1.2867 = 0.70983
3. Only university having a rating with FIVE has a 0.0218 positive correlation with the chance of admission, it is the third important variable that affecting the admission rate. Other universities ranking have no effect on the admission rate.
4. Improving the CGPA, practise more research to gain research experience and receive as many letters of recommendations from professors

Part 2

5. 8
6. Condition: CGPA > 8.53, GRE Score > 318, the predicted result is: accept, the prediction confidence is high, having 204/210 = 0.97 confidence.
7. having a University rating of four and TOEFL score higher than 106, or having a university rating
8. Overall accuracy: (57 + 89)/ (57 + 89 + 16 + 8) = 0.86, misclassification rate: (16+8)/ (57 + 89 + 16 + 8) = 0.14
9. Precision: 89/ (89 + 8) = 0.91, Recall: 89/ (89 + 16) = 0.85, as per the precision and recall, this model has a sightly better performance on precision than recall. For the case of admission prediction, recall is more important than precision, as a higher recall may lead to excellent students missing the chances admitting the programme.

Part 3

10. Cluster 0, it has only 257 customers
11. These customers are: married male, having the second highest average income ($122976.7237) among other clusters, and most of them probably live in a

medium-sized city.

12. There's a positive correlation between customer's income and the size of their hometown: the higher the income, the higher chance that the customer lives in a larger city.
13. Cluster 2, as cluster 2 is the only cluster having single female customers. As in cluster 2, they are more sensitive to the product price (the lowest average income among other cluster groups) and most of them live in small city, we can launch marketing campaign in small cities providing promotion there.

Part 4

14. This is the association rules:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	E	B	0.6	0.8	0.6	1	1.25	0.12	inf	0.5
7	A, E	B	0.6	0.8	0.6	1	1.25	0.12	inf	0.5
9	E	B, A	0.6	0.8	0.6	1	1.25	0.12	inf	0.5
1	A	B	1	0.8	0.8	0.8	1	0	1	0
2	B	A	0.8	1	0.8	1	1	0	inf	0
3	C	A	0.8	1	0.8	1	1	0	inf	0
4	A	C	1	0.8	0.8	0.8	1	0	1	0
5	E	A	0.6	1	0.6	1	1	0	inf	0
6	C, B	A	0.6	1	0.6	1	1	0	inf	0
8	B, E	A	0.6	1	0.6	1	1	0	inf	0

15. Item E, as it appears in all of the association rules having the highest lift, as per Q14, which implies that more customers tend to purchase item E with others.
16. Association rules of E > A and E > B have the same support and confidence, however the lift of E > B is higher than E > A for 0.25, as support of A is 5/5 = 1 while support of B is 4/5 = 0.8, per the lift formula list of (E>B) will be larger than (E>A) for: 1/0.8 – 1/1 = 0.25.

Part 15:

Name	Book1	Book2	Book3	Book4	Book5
John	3	3	2	4	3
David	5	2	4	1	5
Helen	5	2	1	5	4
Max	4	5	1	4	?

17. User-based collaborative filtering:

Find the mean rating by person:

Name	Book1	Book2	Book3	Book4	Book5	**Mean**
John	3	3	2	4	3	**3**
David	5	2	4	1	5	**3.4**
Helen	5	2	1	5	4	**3.4**
Max*	4	5	1	4	?	**3.5**

*For Max rating, I exclude Book5 from calculating the mean

Find the similarities between Max and other people, exclude Book 5:

Formula for Pearson correlation:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

Name	Mean	Similarity with Max
John	3	0.7071067811865476
David	3	-0.42163702135578396
Helen	3.25	0.5134360308102703

Select user with similarity over 0.5 as threshold, hence only John and Helen opinions will be considered.

Predicted Rating for Max on Book5 via user-based collaborative filtering:

$$\hat{R}_{u,i} = \overline{R_u} + \frac{\sum_{v \in U} similarity\,(u,v) * (R_{v,i} - \overline{R_v})}{\sum_{v \in U} similarity\,(u,v)}$$

3.5 + (0.707*(3-3) + 0.513 * (4-3.4))/(0.707 + 0.513) = **3.75**

Item-based collaborative filtering:

Find the similarities between Book 5 and other items.

Formula for Pearson correlation:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2 \sum(y - m_y)^2}}$$

Name	Book1	Book2	Book3	Book4	Book5
John	3	3	2	4	3
David	5	2	4	1	5
Helen	5	2	1	5	4
Max	4	5	1	4	?
Mean	4.25	3	2	3.5	4
Similarity with Book5*	**0.8660254**	**-0.8660254**	**0.65465367**	**-0.7205767**	

*Exclude Max's rate

Select item with top 2 similarity with Book5, hence Book1 and Book2 are selected.

Predicted Rating for Max on Book5 via item-based collaborative filtering:

$$\hat{R}_{u,i} = \frac{\sum_{j \in I}(\text{similarity}(i,j) \cdot R_{u,j})}{\sum_{j \in I} |\text{similarity}(i,j)|}$$

R = (4 * 0.866 + 1 * 0.655)/(0.866 + 0.655) = 2.71

18. The average predicted rating by Max is: (**3.75** + 2.71)/2 = 3.23≈3, therefore Max is probably have an average rating on Book5. Max's average rating of other books is 3.5, which means for Max, book5 will probably have a lower-average rating. I will recommend Book5, but in a lower priority then other books, like book1.

19. I would include Book5 with other books which are interested by Max as a bundle sales, e.g. a bundle sales of Book1 and Book5, as both Books are similar and Book1 is interested by Max. Or after Max finished reading Book1, we initially recommended Book5.

Part 20

20. Doc 1:

    John likes to watch movies. Mary likes movies too.

    Doc 2:

    Mary also likes to watch football games.

    Remove stop words, hence the text will becomes:

    'John likes watch movies Mary likes movies Mary also likes watch football games'

    Tokenized these words, find the frequency of these words in Doc 1 and Doc 2:

    Tokenized doc 1 and doc 2:

	likes	movies	john	watch	mary	also	football	games
0	2	2.0	1.0	1	1	0.0	0.0	0.0
1	1	0.0	0.0	1	1	1.0	1.0	1.0

    TF-IDF vectiorized: $w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$

	also	football	games	john	likes	mary	movies	watch
0	0.000000	0.000000	0.000000	0.352728	0.501938	0.250969	0.705457	0.250969
1	0.470426	0.470426	0.470426	0.000000	0.334712	0.334712	0.000000	0.334712

    TF-IDF vectorization with normalization of l1, so that the sum of TF-IDF in each documents will be equal to 1:

	also	football	games	john	likes	mary	movies	watch
0	0.000000	0.000000	0.000000	0.199006	0.201491	0.100746	0.398012	0.100746
1	0.221301	0.221301	0.221301	0.000000	0.112032	0.112032	0.000000	0.112032

21. Most unique tokens as per the TF-IDF for doc 1:

```
movies      0.398012
likes       0.201491
john        0.199006
mary        0.100746
watch       0.100746
also        0.000000
football    0.000000
games       0.000000
Name: 0, dtype: float64
```

    Movies, likes and john are the most unique items for doc 1.

Most unique items as per the TF-IDF for doc 2:

```
also        0.221301
football    0.221301
games       0.221301
likes       0.112032
mary        0.112032
watch       0.112032
john        0.000000
movies      0.000000
Name: 1, dtype: float64
```

Also, football and games are the most unique items.

```
movies      0.398012
likes       0.201491
john        0.199006
mary        0.100746
watch       0.100746
also        0.000000
football    0.000000
games       0.000000
Name: 0, dtype: float64
```

```
also        0.221301
football    0.221301
games       0.221301
likes       0.112032
mary        0.112032
watch       0.112032
john        0.000000
movies      0.000000
Name: 1, dtype: float64
```

Overall speaking, movies, football and games are the most unique items as per the TF-IDF for both documents.