Homework 3

Deadline: 2020-05-24 23:59

Each of the multiple-choice questions accounts for 5 points

Q1. **Which of the following statements about frequent pattern mining is true?**

C

A) If all the proper subsets of an itemset are frequent, then the itemset itself must also be frequent.
B) If all the proper supersets of an itemset are not frequent, then the itemset itself must also be not frequent.
C) If all the proper subsets of an itemset are not frequent, then the itemset itself must also be not frequent.
D) None of above.

Q2. In general, ensemble methods tend to improve the predictive ability more for ___ methods

A

A) Higher-variance
B) Low-variance
C) Higher-bias
D) Low-bias

Q3. Which of the following plot can help us to determine whether it is worthwhile to invest on collecting more training data?

A

A) Learning curve
B) Fitting plot
C) ROC curve
D) Lift curves

Q4. **We try to compare two models A and B for a binary classification problem (positive/negative), and we find that the two models can achieve the same accuracy score on a balanced training dataset. However, model A tends to falsely predicts that examples to be positive, and model B tends to falsely predict that examples to be negative. Then if we need to apply the model to**

C

imbalanced dataset (i.e., only 10% samples are positive), which model can achieve higher accuracy?

A) Model A
B) Model B
C) Model A or model B (the same)
D) It depends

Q5. Which of the following metrics is used to evaluate the general performance of ranking models?

A

A) ROC curve
B) Precision
C) F1-score
D) True positive rate

Q6. If you want to make fast classifications, and update the model quickly and immediately when some new labelled cases for training are collected, what predictive model is best suited?

A

A) Naïve Bayes
B) Classification tree induction
C) k-Nearest Neighbor
D) Logistic Regression

Q7. In general, if we just need to apply the Bayes model to solve the classification model, i.e., find the best *C* to maximize *P*(*C*|*E*) as follows, then we do NOT need to compute the:

$$p(C = c \mid \mathbf{E}) = \frac{p(\mathbf{E} \mid C = c) \cdot p(C = c)}{p(\mathbf{E})}$$

C

A) $p(C = c)$
B) $p(\mathbf{E} \mid C = c)$
C) $p(\mathbf{E})$
D) None of above

Q8. We now need to apply binary classification models to medical diagnosis, i.e., predict whether a patient has cancer (positive) or not (negative). Then which of the following statement is true if we need to choose from three models?

- **Model-1: Recall score=0.8 and Precision score=0.6;**
- **Model-2: F1 score=0.65;**
- **Model-3: False Positive rate=0.8 and True Positive Rate = 0.6.**

A) Model-1
B) Model-2
C) Model-3
D) All the above

Q9. Which of the following statement is <u>NOT TRUE</u> about two methods of representing text, i.e., A: using bag of n-grams up to 1, B: A: using bag of n-grams up to 5?

A) Features generated by method B will be of higher dimensions
B) Features generated by method B will be less sparse
C) Features generated by method A is a subset of those generate by method B
D) None of above

Q10. Which of these organizations would be the most challenging in applying supervised predictive modelling with regards to the availability of training data?

A) A business school that wants to start a new Master's degree program in Business Analytics, and would like to estimate the likely number of applicants.
B) A grocery store that is trying to identify which of its loyalty-card-carrying customers will spend more than $100 next month.
C) A city government that is trying to predict which district will have the most new shops open up next quarter.
D) An online marketing company that wants to estimate the number of clicks that the ads it serves will receive when shown to a particular population after the ads have been launched for a while.

Q11. Which of the following statement about Apriori algorithm is TRUE?

A) The itemsets with high support will also have high confidence score
B) We can still identify those association rules with minimum support without using Apriori algorithm
C) Apriori algorithm is efficient and hence can help us to handle those super market with millions of inventories
D) None of above

Q12. Which of the following statement about Lift(X→Y) and Confidence(X→Y) is TRUE:

A) $Lift(X \rightarrow Y) \neq Lift(Y \rightarrow X)$
B) $Confidence(X \rightarrow Y) = Confidence(Y \rightarrow X)$
C) If $Lift(X \rightarrow Y_1) = Lift(X \rightarrow Y_2)$, and $Support(Y_1) > Support(Y_2)$, we could know $Confidence(X \rightarrow Y_1) > Confidence(X \rightarrow Y_2)$
D) None of above

Q13. [10 points] A university has extensive dataset on its alumni, including past studies, demographic information by zip code, and past donations. The university is planning to send a deluxe brochure and a donation request to some of the alumni (Total targeting cost is fixed per individual) and has sufficient budget for constructing targeting models and running experiments under the following assumptions:

- **Donation amount may vary.**
- **Alumni may spontaneously make a donation (even when not targeted).**
- **Targeting cost is fixed (such as $15).**
- **Other than the targeting cost, there are no additional costs for alumni who are targeted and decide not to donate.**

Then you are supposed to accomplish following tasks. Note: You just need to write down the correct expected value equations to identify the models that should be constructed. There is no need to further solve/develop the equations.

1) **Use the expected value framework to determine the expected value for difference scenarios [4 points].**

2) **Decompose the problems to sub-tasks that can be solved by different data mining models. List the data mining models (regression or classification models) and the input feature and targeting labels that should be used to develop the model [6 points].**

Q14. **[10 points] We have built a different model for binary classification model to predict whether an example is positive or negative, and we further apply two different datasets to evaluate its performance:**

1) **We firstly apply the model to a balanced data set, and we can get the confusion matrix as follows:**

Actual / Predicted	Positive	Negative
Positive	40	20
Negative	10	30

Meanwhile, the corresponding cost matrix for this task is as follows

Actual / Predicted	Positive	Negative
Positive	100	-10
Negative	-5	0

1) Please compute true positive rate (TPR) and false positive rate (FRP) of our built model and expected revenue of applying our model on this balanced dataset [6 points].

2）Then we apply the same model to an unbalanced training dataset with 20% positive examples and 80% negative examples, please compute the expected revenue of applying our model on this unbalanced dataset (assume that the model would have same true positive rate (TPR) and false positive rate) [4 points].

Q15. [10 points] Given a set of transaction records as follows,

Transaction ID	Items
1	{a, b, d, f}
2	{a, b, c, d, e}
3	{a, b, c, e}
4	{a, b, d}
5	{a, b, c, g}
6	{b, c, e, f}

1) **Please identify the frequent itemsets with minimum support as 0.5 using Apriori algorithm step by step [6 points].**

2) **Further identify the association rules with confidence not lower than 0.8 among above frequent itemsets [4 points].**

Q16. [10 points] Consider the following data set on lung diseases. Your goal is to build a (Naïve/) Bayes classifier that predicts whether a person has Bronchitis or Tuberculosis, given his/her symptoms. And the prior probabilities can simply use the class priors.

X-ray shadow	Dyspnea	Lung inflammation	Disease
Yes	Yes	Yes	Bronchitis
Yes	No	No	Bronchitis
No	Yes	Yes	Bronchitis
Yes	Yes	No	Tuberculosis
No	No	Yes	Tuberculosis
No	Yes	No	Tuberculosis

And we need to predict the lung disease of following 2 new patients: user A, and user B:

User	X-ray shadow	Dyspnea	Lung inflammation
A	No	No	No
B	Yes	Yes	No

1) Please compute the probability of P(Tuberculosis |A) and P(Tuberculosis |B) using Bayes model (Note that NOT Naïve Bayes model) [5 points].

Note: if you cannot perform the calculation of a probability, you can set it to be N/A. And it is possible that it can be equally probable to be either Tuberculosis or Bronchitis.

2) Please determine the most probable lung disease for user A and user B using Naïve Bayes model respectively [5 points].