Assignment 7 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. Imagine, you are working with NPTEL course management team and you want to develop a machine learning algorithm which predicts the number of views on the courses. Your analysis is based on features like the name of instructor, number of courses taught by the same instructor on NPTEL in the past and a few other features. Which of the following evaluation metric would you choose in that case?

   (a) mean square error
   (b) classification accuracy
   (c) F1 score
   (d) precision
   (e) recall

   Sol. (a)
   You can think that the number of views on the course is the continuous target variable which fall under the regression problem. So, mean squared error will be used as an evaluation metrics.

2. Imagine, you are solving a multiclass classification problem with highly imbalanced class. The distribution of the classes is such that, you observed the majority class 99% of the times in the training data. Your model has 99% accuracy after taking the predictions on test data. Which of the following is true in such a case?
   1) Accuracy is not a good metric for imbalanced class problems.
   2) Accuracy is a good metric for imbalanced class problems.
   3) Precision and Recall are good metrics for imbalanced class problems.
   4) Precision and Recall are not good metrics for imbalanced class problems.

   (a) 1 and 2
   (b) 2 and 3
   (c) 1 and 3
   (d) 2 and 4
   (e) 3 and 4
   (f) 1 and 4

   Sol. (c)
   In an imbalanced data set, accuracy should not be used as a measure of performance because 99% (as given) might only be predicting majority class correctly, but our class of interest is also the minority class (1%). Hence, in order to evaluate model performance, we should use Precision, Recall, and F measure to determine class wise performance of the classifier.

3. Imagine you are working on a project which is a binary classification problem. You trained a model on training dataset and get the below confusion matrix on validation dataset.

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Based on the above confusion matrix, choose which option(s) is true among the following?
1) Accuracy is 0.91
2) Misclassification rate is 0.91
3) False positive rate is 0.95
4) True positive rate is 0.95

(a) 1 and 2
(b) 2 and 3
(c) 1 and 3
(d) 2 and 4
(e) 3 and 4
(f) 1 and 4

Sol. (f)
The Accuracy (correct classification) is (50+100)/165 which is nearly equal to 0.91.
The true Positive Rate is how many times you are predicting positive class correctly so true positive rate would be 100/105 = 0.95 also known as "Sensitivity" or "Recall"

4. In identifying frequent itemsets in a transactional database, we find the following to be the frequent 3-itemsets: {B, D, E}, {C, E, F}, {B, C, D}, {A, B, E}, {D, E, F}, {A, C, F}, {A, C, E}, {A, B, C}, {A, C, D}, {C, D, E}, {C, D, F}, {A, D, E}. Which among the following 4-itemsets can possibly be frequent?

(a) {A, B, C, D}
(b) {A, B, D, E}
(c) {A, C, E, F}
(d) {C, D, E, F}

Sol. (d)
By the apriori property, only itemset {C, D, E, F} can possibly be frequent since all of its subsets of size 3 are listed as frequent. The other 4-itemsets cannot be frequent since not all of their subsets of size 3 are frequent. For example, for the first option, the itemset {A, B, D} is not frequent.

5. Consider the following transactional database of 10 transactions.

Transaction ID	Item set
T1	AB
T2	BCD
T3	ACDE
T4	ADE
T5	ABC
T6	ABCD
T7	BA
T8	ABC
T9	ABD
T10	BCE

Making use of the apriori property, find the number of frequent item sets, for a minimum support of 4 (an item set with support greater than or equal to 4 is frequent)

(a) 7
(b) 8
(c) 10
(d) 6

Sol. (b)

6. Consider the following transactional data.

Transaction ID	Items
1	A, B, E
2	B, D
3	B, C
4	A, B, D
5	A, C
6	B, C
7	A, C
8	A, B, C, E
9	A, B, C

Assuming that the minimum support is 2, what is the number of frequent 2-itemsets (i.e., frequent items sets of size 2)?

(a) 2
(b) 4

(c) 6

(d) 8

Sol. (c)
Candidate 1-itemsets:

itemset	support
{A}	6
{B}	7
{C}	6
{D}	2
{E}	2

Frequent 1-itemsets:

itemset	support
{A}	6
{B}	7
{C}	6
{D}	2
{E}	2

Candidate 2-itemsets:

itemset	support
{A, B}	4
{A, C}	4
{A, D}	1
{A, E}	2
{B, C}	4
{B, D}	2
{B, E}	2
{C, D}	0
{C, E}	1
{D, E}	0

Frequent 2-itemsets:

itemset	support
{A, B}	4
{A, C}	4
{A, E}	2
{B, C}	4
{B, D}	2
{B, E}	2

7. For the same data as above, what are the number of candidate 3-itemsets and frequent 3-itemsets respectively?

(a) 1, 1

(b) 2, 2

(c) 2, 1

(d) 3. 2

Sol. (b)
Candidate 3-itemsets:

itemset	support
{A, B, C}	2
{A, B, E}	2

Frequent 3-itemsets:

itemset	support
{A, B, C}	2
{A, B, E}	2

8. Continuing with the same data, how many association rules can be derived from the frequent itemset {A, B, E}? (Note: for a frequent itemset X, consider only rules of the form S $\rightarrow$ (X-S), where S is a non-empty subset of X.)

   (a) 3

   (b) 6

   (c) 7

   (d) 8

   Sol. (b)
   {A} $\rightarrow$ {B, E}

   {B} $\rightarrow$ {A, E}

   {E} $\rightarrow$ {A, B}

   {A, B} $\rightarrow$ {E}

   {A, E} $\rightarrow$ {B}

   {B, E} $\rightarrow$ {A}

9. For the same frequent itemset as mentioned above, which among the following rules have a minimum confidence of 60%?

   (a) $A \wedge B \implies E$

   (b) $A \wedge E \implies B$

   (c) $E \implies A \wedge B$

   (d) $A \implies B \wedge E$

   Sol. (b), (c)
   The confidence values for the above four rules are respectively, 2/4, 2/2, 2/2, and 2/6. Hence, only rules in (b) and (c) have the minimum required confidence.

10. Which of the following statements are true, about frequent itemsets in the context of transactional databases (Note that more than one statement may be correct)

    (a) Every maximal frequent itemset is a closed frequent itemset.

    (b) Every closed frequent itemset is a maximal frequent itemset.

    (c) We can recover all frequent itemsets given all maximal frequent itemsets.

    (d) We can recover the frequencies of all frequent itemsets, given the frequencies of all maximal frequent itemsets.

    Sol. (a), (c)