Politecnico di Torino

Data Science & Machine Learning for engineering applications

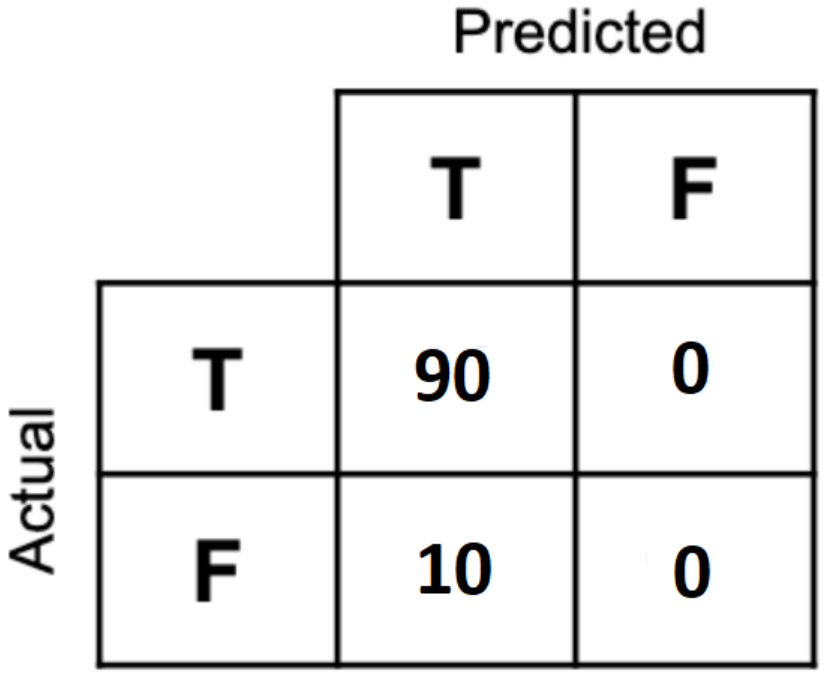
AA 2022-2023

DataBase and Data Mining Group

Prof. Tania Cerquitelli

Question 1

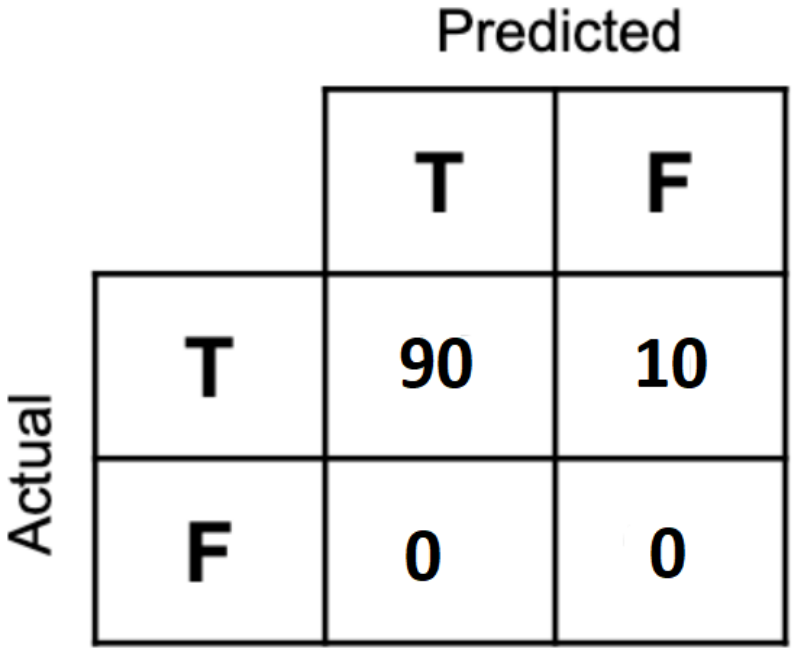
Given the confusion matrix depicted in figure, which statement **is not** correct?



- (a) Precision of class F is 0
- (b) Precision of class T is 90%
- (c) All 100 items are predicted as class T
- (d) Model accuracy is 90%
- (e) Recall of class F is 10%
- (f) Recall of class T is 100%

Question 2

Given the confusion matrix depicted in figure, which statement **is not** correct?



- (a) Recall of class F is 0
- (b) Precision of class T is 100%
- (c) All elements belonging to class T are correctly predicted
- (d) Recall of class T is 90%
- (e) Precision of class F is 0
- (f) The model accuracy is 90%

PoliTo DBMG

Question 3

Which of the following techniques can be used to avoid the "curse of dimensionality"?

- (a) None of the other answers is correct
- (b) Outlier Detection
- (c) Discretization
- (d) Stratified Sampling
- (e) Normalization
- (f) Feature Selection

Question 4

Which of the following techniques can be used to reduce the number of records in a dataset?

- ○ (a) Feature Selection
- ○ (b) None of the other answers is correct
- ○ (c) Outlier Detection
- ○ (d) Stratified Sampling
- ○ (e) Normalization
- ○ (f) Principal Component Analysis

Question 5

A binary decision tree is used as a classification algorithm on a dataset containing 1000 records, 900 belonging to class A and 100 belonging to class B. Which of the following metrics allows an appropriate evaluation of the classifier's performance with regard to class B predictions?

- ○ (a) Accuracy
- ○ (b) Gini Index
- ○ (c) Sum of Squared Error (SSE)
- ○ (d) Mean Squared Error (MSE)
- ○ (e) F-measure
- ○ (f) Average Silhouette Index

Question 6

Which of the following metrics **is not** suitable for evaluating the performance of a classification algorithm?

- (a) Precision
- (b) Confusion matrix
- (c) Silhouette Index
- (d) Accuracy
- (e) F-measure
- (f) Recall

Question 7

The following transactional dataset is given.

B D
A C D E
B D
A E
B C D E
A B C D
A B
A B C E
B C D
A B C E

Write the header table of the corresponding FP-Tree with MinSup>2.

Question 8

The following transactional dataset is given.

B D
A C D E
B D
A E
B C D E
A B C D
A B
A B C E
B C D
A B C E

Write the FP-Tree with MinSup>2. Specifically, report the list of paths characterizing the FP-Tree. For each path specify the sequence of nodes in the form (item, local support).

Question 9

The following transactional dataset is given.

B D
A C D E
B D
A E
B C D E
A B C D
A B
A B C E
B C D
A B C E

Write the node link chain for item E

Question 10

The Apriori principle states that, if an itemset is frequent, then all of its subsets must also be frequent.

You are given the following transactional dataset.

A C E
A B C D
C D
A C D
A B
A B E
B C E
A D E
A B E
A C D E

You need to apply the Apriori algorithm to extract the frequent itemsets with minsup = 2 (an itemset is frequent if it is contained in at least 2 transactions).

During the generation of the candidate itemsets of length 3, after the “prune step” with the application of the “Apriori Principle”, what is the number of candidates extracted for counting their support in the dataset?

Question 10

The Apriori principle states that, if an itemset is frequent, then all of its subsets must also be frequent.

You are given the following transactional dataset.

A C E
A B C D
C D
A C D
A B
A B E
B C E
A D E
A B E
A C D E

Among the set of possible answers,
Select the correct one

- (a) 8
- (b) 9
- (c) 7
- (d) 5
- (e) 6
- (f) 4

You need to apply the Apriori algorithm to extract the frequent itemsets with minsup = 2 (an itemset is frequent if it is contained in at least 2 transactions).

During the generation of the candidate itemsets of length 3, after the “prune step” with the application of the “Apriori Principle”, what is the number of candidates extracted for counting their support in the dataset?