Machine Learning **Sep 6, 2024**

QUESTION SHEET

Question 1 ♣ Considering the list of transactions in the table below. Referring to the association rules, mark the True(correct) statements.

ID	Items
1	{A, B, C}
2	{A, B, C, D}
3	{B, D}
4	{B, D}
5	{A, B}
6	{C, D, E}
7	{B, E}
8	{B, C}
9	{C, D, E}
10	{B, C, D}
11	{B, C, E}
12	{B, D, E}
13	{A, D, E}

■ The confidence of $\{B \Rightarrow D\}$ is 0.5 .

■ If an itemset is NOT frequent, then all its supersets are also NOT frequent.

■ The support of rule $\{\{A, B\} \Rightarrow C\}$ is $\frac{2}{13}$.

D The lift of rule $\{B \Rightarrow D\}$ is less than the lift of $\{B \Rightarrow E\}$.

■ The apriori algorithm helps improve the efficiency of computing strong rules.

F *None of these answers are correct.*

Question 2 ♣ Referring to regression, mark the True(correct) statements.

A If a newly added independent variable improves the model's ability to explain the variance, adjusted R^2 increases more than R^2.

B A slope coefficient with a confidence interval of $[-1, 7]$ can be significant.

■ A regression model showing homoscedasticity means that the residuals have constant variance across all levels of the independent variables.

■ To predict the car price using a decision tree with 20 leaves, we can "at most" output 20 different price values.

E Changing the unit of measurement for one variable in a linear regression model does NOT affect the coefficients of the other variables.

■ To predict the white blood cell count from a Complete Blood Count (CBC) dataset, the Identity activation function (no activation function) should be preferred over the sigmoid for the last neural layer. ($sigmoid(x) = \frac{1}{1+e^{-x}}$)

G *None of these answers are correct.*

Question 3 ♣ Referring to classification, mark the True(correct) statements.

■ A KNN classifier trained on a large dataset typically has a slower inference runtime than a Decision Tree trained on the same dataset.

■ An SVM classifier with an RBF kernel aims to find the optimal separating boundary in a transformed feature space, where it maximizes the margin between classes.

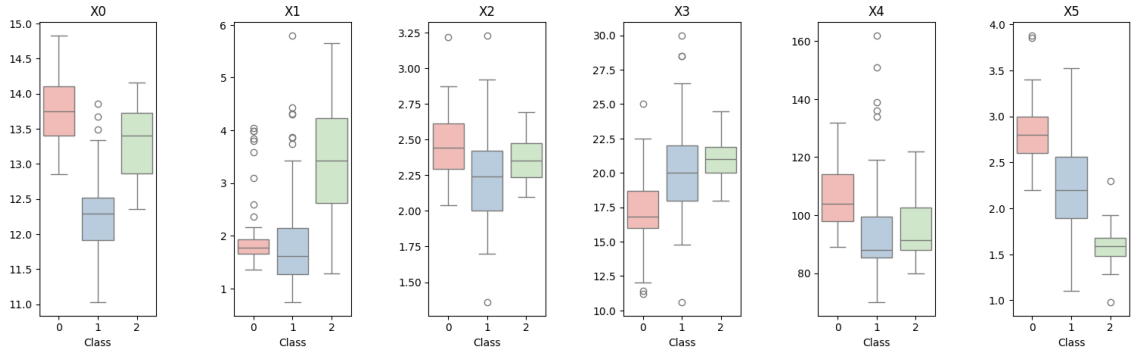
■ If there is no True Positive, F1 score is zero.

D A multilayer perceptron (MLP) classifier with too many hidden neurons is more likely to underfit.

■ A binary classifier that classifies everything as the majority(negative) class the accuracy is equal to the proportion of negative instances over all instances.

F *None of these answers are correct.*

Question 4 ♣ Considering the boxplots shown in the following figure, for a classification task, mark the True(correct) statements.



A Rescaling the features helps improve the performance of a Naive Bayes classifier on this dataset.

■ class '1' of the feature 'X4' has a positive skewness.

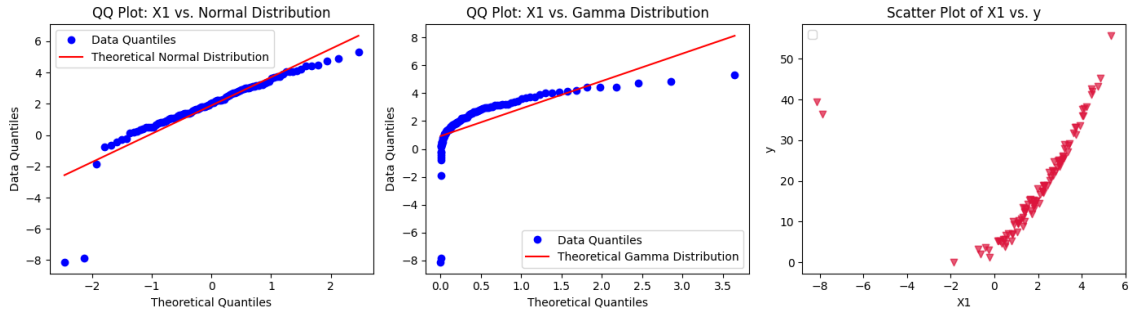
C The feature 'X0' contains at least three outliers associated with class '1'.

■ If the number of observations for each class is 5000, 4000, and 50, upsampling is preferred over downsampling.

■ The feature 'X5' is more likely to be among the most influential features.

F *None of these answers are correct.*

Question 5 ♣ Referring to the figures below, mark the True(correct) statements.



■ Removing the suspicious(extreme) values will improve the Pearson coefficient between $X1$ and y.

B Rescaling the two variables ($X1$, y) will change their Pearson coefficient.

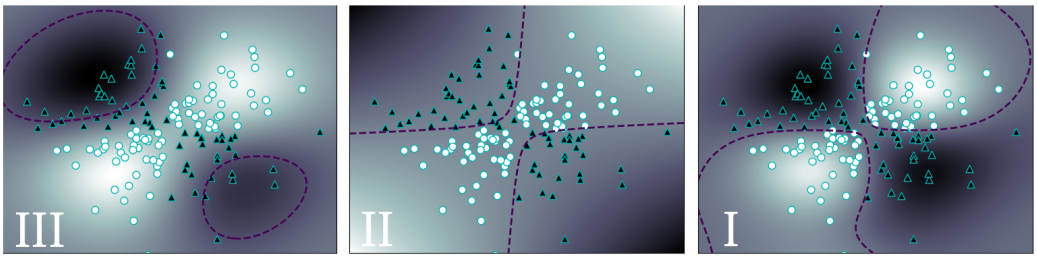
C $X1$ has a right-skewed distribution.

■ The p-value obtained from the KS test when comparing $X1$ to a normal distribution is expected to be higher than the p-value obtained when comparing it to a Gamma distribution.

■ If the kurtosis value of $X1$ is 10.2 and that of y is -0.2, then $X1$ has more observations in its tails compared to y.

F *None of these answers are correct.*

Question 6 ♣ Given a dataset with two numerical explanatory features, the three figures below depict the output of an SVM for a binary classification problem on the dataset. In the training set, circle(white) and triangle(black) points correspond to observations in classes 0 and 1 respectively; while light and dark regions show the prediction values (0 or 1) for new observations according to the three different models.
Mark the True(correct) statements.



A The separating margin size in (III) is narrower than that of (I).

■ Among the three models, it is expected that model (II) gets a higher accuracy value on the test set.

■ The regularization parameter (C) in the model (I) has a high value, giving much importance to the misclassification errors.

■ The kernel used in (I) and (III) is likely to be RBF but with different regularization parameter values.

■ The kernel used in model (II) could be polynomial with degree 2.

F *None of these answers are correct.*

Question 7 ♣ Regarding a piece of code for a classification task with 3 different models, mark the True(correct) statements.

```
from sklearn import metrics
from sklearn.          import
from sklearn.              import
from sklearn.    import
from sklearn.                 import

m1 =                        (n_neighbors = 5)
m2 =                    (fit_intercept = True)
m3 =     (C = 10, epsilon = 0.01, kernel = 'poly', degree = 2)
m4 =              (hidden_layer_sizes = (10, 5), alpha = 0.001, max_iter = 5000)

for m in [m1,m2,m3,m4]:
    m.fit(X_train,y_train)

    y_train_pred = m.predict(X_train)
    y_test_pred = m.predict(X_test)

    MAE_train = metrics.mean_absolute_error(y_train, y_train_pred)
    MAE_test = metrics.mean_absolute_error(y_test, y_test_pred)

    print("MAE train %.3f test %.3f" % (MAE_train, MAE_test))
```

```
MAE train 0.247 test 0.738
MAE train 0.837 test 0.856
MAE train 0.186 test 0.647
MAE train 0.098 test 0.105
```

A m1 can be either a Local Outlier Factor or a KNN.

■ We should decrease C in m3.

■ We should consider increasing the n-neighbors value for m1.

D The m4 model receives 10 input features.

E m2 is the best model among the given models.

F Increasing 'degree' for m3 can reduce MAE on the test set.

G *None of these answers are correct.*

Question 8 ♣ Mark the True(correct) statements.

A A zero correlation between an explanatory feature and the target should be interpreted as the absence of a relation between the variables.

■ The Chebyshev theorem provides an interval around the mean containing a predefined proportion of the observations, that can be used to spot potential outliers.

■ Reducing data dimensions can help prevent overfitting in supervised learning models.

■ GridSearch systematically explores a range of hyperparameters to find the best combination for model performance.

E In the case of data incompleteness, it is always preferable to consider elimination techniques to prevent the injection of noise into the dataset.

F *None of these answers are correct.*

Question 9 ♣ Referring to the clustering problem, mark the True(correct) statements.

- ■ The silhouette coefficient of the observations can take a negative value if it inherently belongs to another cluster.
- ■ In clustering using complete linkage, when you cut the dendrogram at a height of h, you ensure that the largest distance within any cluster does not exceed h.
- ■ When the variables are correlated or have different variances, the Mahalanobis distance is preferred over Minkowski.
- ■ K-medoids is more robust to noise and outliers compared to K-means, as medoids are less influenced by extreme values.
- E for large datasets, divisive algorithms are preferred over agglomerative.
- F *None of these answers are correct.*

Question 10 In the text box of this question, you can add any comment related to the test (withdraw, clarification, etc).

Question 11 ♣ Black out the answer to this question (**in the answer sheet!**) to withdraw from the exam.

■

Corrected

ANSWER SHEET

- **Make sure you read the instructions carefully.**
- Total time: 50 minutes.
- Usage of texts, notes, or digital devices is forbidden.
- Questions can have multiple correct answers.
- Wrong answers are attributed a negative score (half the score of a correct answer).
- Completely blacken the box beside a correct answer **in the answer sheet ONLY**. Answers indicated in other sheets will be ignored. Use a white-out marker to correct and unmark previously marked boxes.

  DOs: ☐ ■ **DON'Ts**: ☒ ☒ ☑ **Absolute DON'Ts**: ■ (NO)

- No questions are allowed during the exam. Comments can be included in the last box. If your comment refers to a question, make sure you reference the question (for example, "In question 1, answer B, ...").
- It is your responsibility to write in a clear and understandable way.
- **DO NOT** black out the boxes in the top right corner of open text questions.
- **To withdraw** from the exam, completely blacken out the box associated to Question 11 in the answer sheet.
- Please encode your **codice persona** (8 digits) below, blacking out one digit per column. For instance, if your codice persona is 12345678, the "1" digit must be blacked out in the leftmost column, and the "8" digit must be blacked out in the rightmost column.

0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9

- Write your first and last names in the box below.

Last name: .....

First name: .....

Answers:

Question 1: ■ ■ ■ D ■ F

Question 2: A B ■ ■ E ■ G

Question 3: ■ ■ ■ D ■ F

Question 4: A ■ C ■ ■ F

Question 5: ■ B C ■ ■ F

Question 6: A ■ ■ ■ ■ F

Question 7: A ■ ■ D E F G

Question 8: A ■ ■ ■ E F

Question 9: ■ ■ ■ ■ E F

Question 10: ■

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

Question 11: ■