Midterm Test

1. This is an open-book test, which accounts for 20% of your total score.
2. Total Score is 100.
   - Part I Binary and Multiple Choices Questions (36%)
   - Part II Descriptive Questions (20%)
   - Part III Empirical Questions (44%)
3. Submission format.
   - Please organize all answers in one file (PDF, HTML or WORD) for submission.
   - For answers with Python, please report both the codes and results clearly.
   - For answers with manual calculation, please report all calculation steps clearly.

Part I. Binary and Multiple Choices Questions [36 Points]

There are 23 questions in total, with only one correct answer for each question. Please organize your answers together with question number clearly.

(A) Binary Choice Questions [10 points @ 1 points each]

Determine which is the best approach (A or B) for tasks in Question 1-5.

A. *supervised learning* B. *unsupervised learning*

1	Use outstanding balance and age to predict whether customers will default on the loan.
2	Separate customers into different groups based on their purchasing records and gender.
3	Use sales revenue from the last quarter to predict sales revenue in next quarter.
4	Group all staff into different types based on their job performance.
5	Separate credit card transactions into different types based on transaction time and location.

Determine whether the statements in Question 6 - 10 is

A. *True* B. *False*

6	Decision Tree models are non-parametric supervised learning method.
7	Adding the complexity to a model usually increase its performance on test data.
8	Support-Vector Machines (SVMs) can model non-linear relationships with kernel functions.
9	Logistic regression is an unregularized model, it doesn't control the coefficient size.
10	For SVMs, a smaller C value means stronger regularization on the coefficients.

(B) Multiple Choice Questions [26 points @ 2 points each]

11. In a Decision Tree model which predicts the risk level of customers, a leaf node contains 50 training instances with different values of **Risk Level**, and the corresponding frequency as below.

Risk Level	Frequency
High Risk	30
Medium Risk	10
Low Risk	10

For an instance in this leaf node, what is the estimated "High Risk" probability for it if we adopt the Laplace Method?

A. 0.2 B. 0.596 C. 0.6 D. 0.608

12. Common metrics used to evaluate a Linear Regression model are_____.

A. *R-Squared*: the lower the better & *Mean Squared Error*: the higher the better
B. *R-Squared*: the lower the better & *Mean Squared Error*: the lower the better
C. *R-Squared*: the higher the better & *Mean Squared Error*: the higher the better
D. *R-Squared*: the higher the better & *Mean Squared Error*: the lower the better

13. If a Logistic Regression model is perfectly accurate on the training data (i.e., train error rate is zero), then _____.

A. Test error rate is also always zero
B. Test error rate tends be smaller than Train error rate
C. Test error rate is equal to Train error rate
D. Test error rate tends to be non-zero

14. Which of the following statements about the fitting graph is NOT correct?

A. It shows the training performance vs. model complexity.
B. It shows the generalization performance vs. model complexity.
C. It shows the generalization performance vs. training data size.
D. It helps decide the complexity level for a model in order to achieve the best generalization performance.

15. Which of the following statements about model complexity is NOT true?
    A. Complex model tends to take less time to be trained than simpler models
    B. Complex model tends to have worse generalization performance than simpler models.
    C. Complex model tends to overfit more than simpler models.
    D. Complex model tends to be less interpretable than simpler models.

16. When training a decision tree model with the ***DecisionTreeClassifier*** function from ***scikit-learn***, which of the following parameters can NOT be used to control tree size?
    A. max_depth
    B. max_leaf_nodes
    C. C
    D. min_samples_leaf

17. Logistic Regression is a _____ model used to predict a _____outcome.
    A. linear, continuous
    B. linear, discrete
    C. nonlinear, continuous
    D. nonlinear, discrete

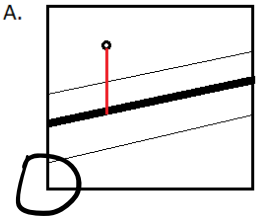
18. Which of the following statements about Logistic Regression is NOT true?
    A. It takes a discrete target variable.
    B. Training a logistic regression using the scikit-learn package requires numeric features.
    C. The linear values (i.e., confidence scores) represent the odds of positive class membership.
    D. The linear values (i.e., confidence scores) represent the log-odds of positive class membership.

19. Which of the following statements is NOT true about Support Vector Machine for classification ?
    A. SVM allows instances to be in the incorrect side of its margin line or even the hyperplane.
    B. SVM finds the best hyperplane by minimizing the sum of hinge loss on training instances.
    C. Instances violating their margin line, but not the hyperplane, will be classified correctly.
    D. Instances violating the hyperplane will receive a smaller hinge loss than those violating their margin line only.

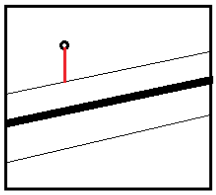
20. In the following figure, the bold line in the middle represents the separating hyperplane and the two thin lines represent the margin lines of a SVM model; the circle is an training instance and the line connecting the circle to either the hyperplane or margin line is the distance.

For instance x_i, which distance does $\boldsymbol{y_i} * (\boldsymbol{w_0} + \boldsymbol{w_1 x_1} + \boldsymbol{w_2 x_2})$ measure given $\sum_{j=1}^{2} w_j^2 = 1$? Choose a proper distance from the following options. (*Note:* y_i *= 1 when instance's actual class is positive, otherwise* y_i *= -1.*)

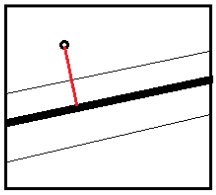
A.



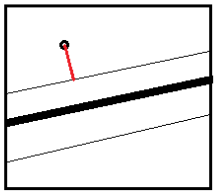
B.



C.



D.



You may need to type and run the following Python codes to answer Question 21-23.

```
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression

np.random.seed(2025)
x = np.random.normal(loc=3, scale=1, size=100)
e = np.random.normal(loc=1, scale=1, size=100)
y = 1 + 2 * x + e

df = pd.DataFrame({"X":x, "y":y})

model = LinearRegression().fit(df[['X']], df['y'])
```

21. Which of the following statements correctly describe variable $\boldsymbol{x}$?
    A. $\boldsymbol{x}$ is a matrix with 2 rows and 1 column. Each value in $\boldsymbol{x}$ is equal to 100.
    B. $\boldsymbol{x}$ is a 1-dimensional array with 100 numbers randomly drawn from a normal distribution with mean as 1 and standard deviation as 2.
    C. $\boldsymbol{x}$ is a 1-dimensional array with 100 numbers randomly drawn from a normal distribution with mean as 3 and standard deviation as 1.
    D. For the 100 values in variable $\boldsymbol{x}$, their mean is 3 and standard deviation should be 1.

22. Which of the following statements is NOT correct regarding the R^2 of a regression *model*?
    A. R^2 score measures the proportion of sum of squared residual to total sum of squares in y.

B. R^2 score measures the proportion of variation in y explained by this model.

C. R^2 score can be obtained with the **score** method associated with the model.

D. R^2 score can be obtained with the ***r2_score*** function from **sklearn.metrics** module.

23. Which of the following statements is NOT correct regarding the Mean Squared Error (MSE)?

A. MSE shows the average squared residual (error) of a regression model.

B. The residual is the difference between the predicted and actual value for the target y.

C. The smaller MSE is on the training data, the better the model fit to the training data.

D. The smaller MSE is on the test data, the worse the model's generalization performance is.

Part II. Descriptive Questions [20 points@ 10 points each]

1. **Model Training**: What is Logistic Regression? Take binary classification as an example, describe how the algorithm find the optimal parameter values and estimate class probabilities? (max: 200 words)

2. **Model Evaluation**: What is overfitting and why it is harmful? How to avoid overfitting in a systematic way? (max: 200 words)

Part III. Empirical Questions [44 Points]

1. Logistic Regression [8 points @ 4 points each]

A sample of 580 customers whose ***decision*** on purchasing a house and ***income*** was collected. A researcher uses ***income*** to predict ***customers'*** ***decision*** ($y = 1$ if the customer will buy, $y = 0$ if not) with a logistic regression model. The model returns the optimized parameters $w_0 = -1.5$ and $w_1 = 0.15$, i.e. $f(x) = -1.5 + 0.15 * income$.

Please answer the following questions with Python or manual calculation.

A. What is the probability that a customer will buy a house if his/her ***income*** is 20? Round the result to 2 decimal places.

B. What is the predicted decision for this customer?

2. Data Generation and Visualization [16 points @ 4 points each]

Use Python to complete the following tasks.

A. Create 100 numbers evenly spaced between in the range [0, 10), with step size as 0.1. Save them in a variable named $\boldsymbol{x1}$. (*Note: 10 shall not be included.*)

B. Apply natural exponential transformation to $\boldsymbol{x1}$ values, save the transformed values as variable $\boldsymbol{x2}$.

C. Save the two variables in a data frame named $\boldsymbol{df}$, check the mean value for each variable.

D. Visualize the relationship between $\boldsymbol{x1}$ and $\boldsymbol{x2}$ with a scatterplot, with $\boldsymbol{x1}$ values displayed on the x axis and $\boldsymbol{x2}$ values on the y axis. Set the point color of as 'gold', add axis labels and plot title.

3. SVM: Programming [20 points]

The ***Wisconsin Breast Cancer*** dataset from **scikit-learn** package contains 30 features and 1 target variable for 569 patients. The features records different characteristics of a patient's tumour, while the target variable refers to the diagnosis: 0 means benign, 1 means malignant.

```
from sklearn.datasets import load_breast_cancer

X, y = load_breast_cancer(return_X_y = True, as_frame = True)

X = X[['mean radius', 'mean texture', 'mean perimeter', 'mean area','mean smoothness']]

display(X.shape, y.shape)
```

Please type and run the above codes to obtain data, select 5 features (***X***) and the target variable (***y***). Use the selected data to answer the following questions with Python.

A. **[3 points]** Split the data into train (80%) and test (20%) sets, with *random_state* = 2025. How many instances are there in the train and test set respectively?

B. **[5 points]** Scale the features with the ***MinMaxScaler***. After scaling, what are the minimum and maximum values for test features respectively? Why none of the minimum values is 0 and none of the maximum values is 1?

C. **[5 points]** With the scaled data from step (B), train two linear SVM models using the ***SVC*** function from the ***scikit-learn*** package, set C value as 0.1 and 100 respectively. Check the two models' performance on both train and test set, round the results to 2 decimal places. Which model is better and why?

D. **[7 points]** Among the ten C values [0.001, 0.01, 0.1, 1, 10, 50, 100, 150, 200, 500], which is the best for the Linear SVM model on this dataset? Please use the scaled training data from step (B) and Grid Search with 5-fold Cross Validation to find the best C value. Answer the following questions: (1) what is the best C value found? (2) what is the average generalization performance of the best C value in the cross-validation process? (3) what is the generalization performance of the best model (refitted on the training data with the best C value found in cross-validation)? Round results to 2 decimal places.