Enabling People
Education for Future Generations

AI Course

Chapter 5. Quiz

For students

Samsung Innovation Campus

1. Which of the following statements about supervised learning is incorrect?

   ① Supervised learner is to make predictions correctly with given data from training data.

   ② Precision is the ratio of what the model classifies as true to what it actually classifies as true. This can be called the answer rate.

   ③ Recall ratio is the ratio of what model predicts to be true out of what is actually true.

   ④ Recall ratio is calculated as $\frac{tp}{tp+fp}$. (where tp is true positive and fp is false positive.)

ANSWER: 4
Option 4 is incorrect. Recall is calculated as tp/(tp+fn), not tp/(tp+fp). The formula tp/(tp+fp) is actually the formula for Precision. Recall measures the ratio of true positives among all actual positives (tp+fn).

2. What is the method of estimating regression coefficients in regression analysis?

ANSWER: Ordinary Least Squares (OLS) / Least Squares Method
The Ordinary Least Squares (OLS) method is the standard approach for estimating regression coefficients. It minimizes the sum of squared residuals (differences between observed and predicted values).

3. Which of the following statements about overfitting is incorrect?

   ① A characteristic of a pattern found in a specific dataset that exists only in that dataset, or a characteristic of a pattern that does not generalize and exists only in a specific data is called overfitting.

   ② The extraction methods used to solve the overfitting problem include the holdout method, the cross-validation method, the bootstrap method, etc.

   ③ In the holdout method, cross-validation is performed by dividing the training data for model learning and construction and the verification data for performance evaluation, and the results of the verification data are used only for performance measurement without affecting the model.

   ④ The cross-validation method is a method to reselect training data repeatedly and based on restoration extraction, and it is suitable when the total amount of data is not large.

ANSWER: 4
Option 4 is incorrect. The statement describes the bootstrap method, not cross-validation. Cross-validation divides data into k-folds and trains k times using different validation sets WITHOUT replacement. Bootstrap uses sampling WITH replacement (restoration extraction).

4. Which of the following about the decision tree model is incorrect?

   ① Decision trees are used for corporate bankruptcy prediction, stock price cap prediction, exchange rate prediction, and economic outlook prediction.

   ② CHAID(Chi-squared Automatic Interaction Detection) algorithm is an algorithm that performs separation using chi-squared or F-test.

   ③ CART(Classification and Regression Trees) algorithm is an algorithm that performs separation using the Gini index, and 100 represents a perfectly pure node between 0 and 100.

   ④ The entropy index of the C4.5 algorithm can be obtained by using the likelihood ratio test statistic in the polynomial distribution and subtracting the entropy of the child node from the entropy of the parent node.

ANSWER: 3
Option 3 is incorrect. The Gini index ranges from 0 to 1 (or 0 to 0.5 for binary classification), not 0 to 100. A Gini index of 0 represents a perfectly pure node (all samples belong to one class), while higher values indicate more impurity.

5. When I trained a SVM classifier using the RBF kernel, it seems to underfit the training set. Should we increase gamma or decrease gamma? What about C?

ANSWER: Increase both gamma and C
To address underfitting: (1) INCREASE gamma - makes the decision boundary more complex and fit training data better. Higher gamma means each training example has more localized influence. (2) INCREASE C - reduces regularization, allowing the model to fit the training data more closely by penalizing misclassifications more heavily.

6. Use the KNN algorithm to classify the data in the code below and calculate the accuracy, call, and precision. Set the adjacency value to 5.

```
import numpy as np
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.neighbors import KNeighborsClassifier
from sklearn import metrics, preprocessing

# Load data.
data = load_breast_cancer()
X = data['data']
Y = 1 - data['target'] # 0 = 'benign' and 1 = malignant.

# Split the dataset into training and testing.
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.4, random_state=1234)
```

ANSWER: Code solution provided below

```
# Create and train KNN classifier with k=5
knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(X_train, Y_train)
Y_pred = knn.predict(X_test)
# Calculate metrics
accuracy = metrics.accuracy_score(Y_test, Y_pred)
recall = metrics.recall_score(Y_test, Y_pred)
precision = metrics.precision_score(Y_test, Y_pred)
print(f'Accuracy: {accuracy:.4f}')
print(f'Recall: {recall:.4f}')
print(f'Precision: {precision:.4f}')
```