Introduction to Supervised Learning with scikit-learn

Accuracy

A spam recognition classifier is described by the following confusion matrix:

TP, TN, FP, FN = 4, 91, 1, 4

Compute the accuracy and insert the answer in the box below.

Ans : 0.95

Accuracy

A spam recogition classifier is described by the following confusion matrix:

TP, TN, FP, FN = 0, 95, 5, 0

Compute the accuracy and insert the answer in the box below.

Ans : 0.95

Precision

Compute the precision based on the following results:

TP = 114 FP = 14

Ans : 0.89

Recall

Compute the recall based on the following results:

TP = 114 FN = 0

Ans : 1.00

F1-SCORE

Compute the F1- score based on the following results:

TP, FP, FN = 2.00, 1.00, 90.00

Ans : 0.04

Fit a logistic regression model using the given input and output patterns X and Y, respectively. Once fitted, determine the precision of this fitted logistic regression model using the test dataset.

X\_train = [[4,2,1],[3,4,6],[5,6,7],[8,9,7]]

y\_train = [1,2,1,2]

X\_test = [[4,3,1],[2,4,3],[5,6,1],[5,9,9]]

y\_test = [1,2,2,2]

Use random\_state=0 in the model and average='weighted'to calculate the precision. round to two decimal places the result.

Ans : 0.83

Fit a Decision Tree model using the input patterns X and the corresponding output patterns Y. After fitting, evaluate the recall of the fitted Decision Tree model using the test dataset.

X\_train = [[4,2,1],[3,4,6],[5,6,7],[8,9,7]]

y\_train = [1,2,1,2]

X\_test = [[4,3,1],[2,4,3],[5,6,1],[5,9,9]]

y\_test = [1,2,2,2]

Use random\_state=0 in the model and average='weighted'to calculate the recall. round to two decimal places the result.

Ans : 0.75

Fit a Decision Tree and a Logistic Regression model using the input patterns X and the corresponding output patterns Y. After fitting, compare the performance of both models in terms of F1-score and determine which one presents the best results.

Decision tree

Logistic Regression

Both models presents similar values of f1-score

Ans : Decision Tree

Precision and Recall

Let’s say we have a machine that classifies if a fruit is an apple or not.

Precision measures how many of our classified oranges were actually oranges.

Recall measures how many of our classified oranges were actually apples.

Recall measures how many apples we might have missed in the entire sample of fruit.

Precision will measure the amount of misclassified oranges as apples (False Positives) and the amount of apples not correctly classified as apples (False Negatives).

Ans : A and C

Evaluation Metrics for Classification

Accuracy

A spam recognition classifier is described by the following confusion matrix:

TP, TN, FP, FN = 4, 91, 1, 4

Compute the accuracy and insert the answer in the box below.

Ans : 0.95

Accuracy

A spam recogition classifier is described by the following confusion matrix:

TP, TN, FP, FN = 0, 95, 5, 0

Compute the accuracy and insert the answer in the box below.

Ans : 0.95

Precision

Compute the precision based on the following results:

TP = 114 FP = 14

Ans : 0.89

Recall

Compute the recall based on the following results:

TP = 114 FN = 0

Ans : 1.00

F1-SCORE

Compute the F1- score based on the following results:

TP, FP, FN = 2.00, 1.00, 90.00

Ans : 0.04

Fit a logistic regression model using the given input and output patterns X and Y, respectively. Once fitted, determine the precision of this fitted logistic regression model using the test dataset.

X\_train = [[4,2,1],[3,4,6],[5,6,7],[8,9,7]]

y\_train = [1,2,1,2]

X\_test = [[4,3,1],[2,4,3],[5,6,1],[5,9,9]]

y\_test = [1,2,2,2]

Use random\_state=0 in the model and average='weighted'to calculate the precision. round to two decimal places the result.

Ans : 0.83

Fit a Decision Tree model using the input patterns X and the corresponding output patterns Y. After fitting, evaluate the recall of the fitted Decision Tree model using the test dataset.

X\_train = [[4,2,1],[3,4,6],[5,6,7],[8,9,7]]

y\_train = [1,2,1,2]

X\_test = [[4,3,1],[2,4,3],[5,6,1],[5,9,9]]

y\_test = [1,2,2,2]

Use random\_state=0 in the model and average='weighted'to calculate the recall. round to two decimal places the result.

Ans : 0.75

Fit a Decision Tree and a Logistic Regression model using the input patterns X and the corresponding output patterns Y. After fitting, compare the performance of both models in terms of F1-score and determine which one presents the best results.

Decision tree

Logistic Regression

Boths models presents similar values of f1-score

Ans : Decision Tree

Precision and Recall

Let’s say we have a machine that classifies if a fruit is an apple or not.

Precision measures how many of our classified oranges were actually oranges.

Recall measures how many of our classified oranges were actually apples.

Recall measures how many apples we might have missed in the entire sample of fruit.

Precision will measure the amount of misclassified oranges as apples (False Positives) and the amount of apples not correctly classified as apples (False Negatives).

Ans : A and C

Ice Cream Sales

Select whether this problem is a regression or a classification problem.

Regression

Classification

Ans : Regression

Identifying Targets and Features for Regression Analysis

Q . Which of the following scenarios represents a suitable regression target variable?

Predicting the type of flower species based on petal length and width.

Estimating the time it takes to complete a marathon race.

Classifying whether an email is spam or not.

Identifying whether a patient has a certain medical condition.

Ans : Estimating the time it takes to complete a marathon race.

Q. In which of the following situations is regression the most appropriate technique?

Determining the sentiment (positive/negative) of a movie review.

Forecasting the price of a house based on its features.

Categorizing images of animals into different species.

Predicting whether a customer will buy a product.

Ans : Forecasting the price of a house based on its features.

Introduction to Scikit-Learn Lab

Q1 . True or False: In Scikit-Learn, each machine learning model is implemented as a class, ensuring consistent interfaces and making it easy to work with different models using similar methods and attributes.

Ans : True

Q2 . True or False: To use a specific machine learning model in Scikit-Learn, you need to create an instance of its corresponding class using the model's constructor, which initializes the model with default or user-specified settings.

Ans : True

Introduction to Using a Pre-Trained Model in Scikit-Learn

In the lab, what is the main objective of using a pre-trained machine learning model?

To visualize the decision boundaries of the model.

To fine-tune the model on a different set of hyperparameters.

To make predictions on new data efficiently without retraining the model.

To train a model from scratch on a new dataset.

Ans : To make predictions on new data efficiently without retraining the model.

Which library is used to save and load the pre-trained model in the lab?

Scikit-Learn

Joblib

NumPy

Matplotlib

Ans : Joblib

What type of model is used in the lab for the linear regression example?

Decision Tree

Logistic Regression

Linear Regression

Ans : Linear Regression

Model Parameters and Hyperparameters Overview

Which of the following is NOT a machine learning model used in the exercise?

Support Vector Machine (SVM)

Random Forest

Linear Regression

Neural Network

Ans : Linear Regression

How can changing hyperparameter values impact a model's performance?

It can only improve the model's performance.

It has no effect on the model's performance.

It only affects the model's visualization.

It can lead to both improvements and deteriorations in performance.

Ans : It can lead to both improvements and deteriorations in performance.

True or False: The accuracy of the SVM with linear kernel model decreases with low values of C.

False

True

Ans : True

True or False: The accuracy of the Random Forest model decreases with low values of max\_depth (with max\_depth=2 and 5).

False

True

Ans : False

Understanding Target and Features for classification

Q . What is the role of feature variables in a classification problem?

To provide additional context to the target variable

To assess the performance of the model

To determine the accuracy of the model

To classify the target variable into different categories

Ans : To classify the target variable into different categories

Q . Which of the following best defines the target variable in a classification problem?

The variable used for feature engineering

The variable used for making predictions

The variable used for normalization

The variable being predicted

Ans : The variable being predicted

Q . In a binary classification problem to predict whether a customer will churn or not, what would be the target variable?

Customer ID

Customer income

Customer age

Likelihood of churn

Ans : Likelihood of churn

Q . Identify the target variable for this problem

BMI

currentSmoker

TenYearCHD

diabetes

Ans : TenYearCHD

Q . Based on previous tasks identify if this problem is a binary of multiclass classification problem

Binary

Multiclas

Ans : Binary

Q . You are given a dataset containing information about customers of an e-commerce website. The goal is to predict whether a customer will make a purchase or not based on their characteristics. Identify the target variable and the features variables in this scenario.

Target variable: Purchase id; Features variables: Customer age, Time spent on the website, Number of products viewed.

Target variable: Customer will make a purchase; Features variables: Customer age, Purchase amount, Time spent on the website, Number of products viewed.

Target variable: Customer age; Features variables: Purchase amount, Time spent on the website, Number of products viewed.

Target variable: Purchase amount; Features variables: Customer age, Time spent on the website, Number of products viewed.

Ans : Target variable: Customer will make a purchase; Features variables: Customer age, Purchase amount, Time spent on the website, Number of products viewed.

Why hyperparameter tuning is important

In the hyperparameter tuning lab, which evaluation metric is used to measure the performance of the logistic regression model?

Precision

Recall

F1-score

Accuracy

Ans : Accuracy

What is the purpose of hyperparameter tuning in machine learning?

To select the best features for the model.

To adjust the model's internal parameters during training.

To find the best combination of hyperparameters for the model.

Ans : To find the best combination of hyperparameters for the model.

what does the plot in final step of the lab indicate?

The testing accuracy of the logistic regression model.

The impact of hyperparameter tuning on the model's computational resources.

The relationship between regularization strength and feature importance.

The training accuracy of the logistic regression model.

Ans : The testing accuracy of the logistic regression model.

feature\_engineering\_simulated\_dataset

Handling Missing Values

Identify and handle missing values in the dataset df\_encoded. This can be done by filling missing values with mean, median, or mode, or by removing rows or columns with missing values.

Select the correct code to replace the missing values for the mean value.

df\_filled = df\_encoded.fillna(df\_encoded.median())

df\_filled = df\_encoded.fill(df\_encoded.mean())

df\_filled = df.fillna(df\_encoded.mean())

df\_filled = df\_encoded.fillna(df\_encoded.mean())

Ans : df\_filled = df\_encoded.fillna(df\_encoded.mean())