

Introduction to Course Projects of Statistical Learning Theories and Applications

1. Introduction

There are four projects for this course: (a) Regression Model Project; (b) SVM (Support Vector Machine) Project; (c) Multi-layer Perceptron Project; (d) Adaboost Project.

2. Submission Deadline and Method

All Projects' Submission Deadline is before 23:59PM UTC+8, Dec 12, 2022.

Please zip the required documents according to the specific requirements of each project as "X+Y.RAR" to TA's email address 17360075977@163.com, with X replaced by the Student Name and Y replaced by the Student ID.

3. Project Specification

3.1 Regression Model Project

Requirement: Please use a regression model to model the data in the "regression" directory. Note that you need to determine the complexity and solving method of the regression model and implement the model by programming by yourself. The existing programming models such as regression interfaces, functions or codes from the third-party are not allowed.

There are three data files in the data set, all in csv format, in which the training set has 10000 samples, and the validation set and test set have 1000 samples. The training set and validation set contain label information to train and validate the model. The test dataset is used to evaluate the learning performance of the model. Please write the prediction results of the test set into the y column of "regression_test.csv" file, when submitting the testing results. The regression prediction is a real number for each feature.

Programming Language: Not specified. Matlab or Python is recommended.

Grading Criteria: MSE (Mean square error) performance of the "regression_test.csv" file, and readability of source files.

Submission Content: Source code files and "regression_test.csv".

3.2 SVM (Support Vector Machine) Project

Requirement: Please use a SVM to model the data in the "classification" directory. Note

that you could use an off-the-shelf model (interface, function and code of third party), but the hyper-parameters of model, such as penalty factor C and kernel function, should be set by yourself.

There are three data files in the data set, all in csv format, in which the training set has 10000 samples, and the validation set and test set have 1000 samples. The training set and validation set contain label information to train and validate the model. The test dataset is used to evaluate the learning performance of the model. Please write the prediction results of the test set into the y column of “svm_classification_test.csv” file, when submitting the testing results. The classification result is 0 or 1 for each feature.

Programming Language: Not specified. Matlab or Python is recommended.

Grading Criteria: Accuracy performance of the “svm_classification_test.csv” file, and readability of source files.

Submission Content: Source code files and “svm_classification_test.csv”.

3.3 Multi-layer Perceptron Project

Requirement: Please use a multi-layer perceptron model to model the data in the “classification” directory. Note that you need to determine and implement the model by programming by yourself. The existing programming models such as multi-layer perceptron interfaces, functions or codes from the third-party are not allowed.

There are three data files in the data set, all in csv format, in which the training set has 10000 samples, and the validation set and test set have 1000 samples. The training set and validation set contain label information to train and validate the model. The test dataset is used to evaluate the learning performance of the model. Please write the prediction results of the test set into the y column of “mlp_classification_test.csv” file, when submitting the testing results. The classification result is 0 or 1 for each feature.

Programming Language: Not specified. Matlab or Python is recommended.

Grading Criteria: Accuracy performance of the “mlp_classification_test.csv” file, and readability of source files

Submission Content: Source code files and “mlp_classification_test.csv”.

3.4 Adaboost Project

Requirement: Please use a Adaboost to model the data in the " classification " directory. Note that you could use an off-the-shelf model (interface, function or code) as the base classifier, but you should realize a whole Adaboost model by yourself.

There are three data files in the data set, all in csv format, in which the training set has

10000 samples, and the validation set and test set have 1000 samples. The training set and validation set contain label information to train and validate the model. The test dataset is used to evaluate the learning performance of the model. Please write the prediction results of the test set into the y column of “adaboost_classification_test.csv” file, when submitting the testing results. The classification result is 0 or 1 for each feature.

Programming Language: Not specified. Matlab or Python is recommended.

Grading Criteria: Accuracy performance of the “adaboost_classification_test.csv” file, and readability of source files.

Submission Content: Source code files and “adaboost_classification_test.csv”.