

Special Crafted by
Bernardus Ari
Kuncoro

Updated
Q4.2019

Practice Case 05 & 06

Supervised Learning (Basic and Advanced)

Unsupervised Learning

Supervised Learning

Given the **train.csv** and **test.csv**, perform the following actions.

1. Create classifiers from train.csv data, in order to predict **att10** as label based on the rest of the attributes. Use at least 5 algorithms (k-NN, Decision Tree, Logistic Regression, Voting, Averaging, Bagging, Random Forest, Averaging, Voting, AdaBoost, XGBoost, LightGBM, CatBoost, or Stacking). Use AUC for your model evaluation performance.

Submit your .ipynb file that contains your coding process in creating classifiers.

2. Choose the best Classifier based on highest AUC and use it for predicting the test.csv data.

*Submit your .csv file that contains the prediction of test.csv data. Sample of prediction can be found in **sample_prediction.csv** file.*

Unsupervised Learning

1. Select att1 and att2 from train.csv and test.csv data. This step will result a new data frame or a table with size of 10,000 rows and 2 columns. Perform clustering using at least 3 algorithms (K-Means Clustering, Hierarchical Clustering, and DBSCAN) to cluster that new data frame.

Submit your .ipynb file that contain the process of your clustering!

2. Assume that the dataset you are working is the employee data, whether

att1 is satisfaction level (the higher the satisfaction level, the more satisfy a person does in his employment)

att2 is evaluation result (the higher the evaluation result, the better/harder a person works in the company).

Can you name the clusters that you have just made?

Rules

1. All submission must be submitted **before 11 November 2019, 11:59pm**
2. Only for supervised learning practice case, you can perform feature engineering, feature selection, undersampling, and oversampling techniques to achieve high AUC.
3. For unsupervised learning, you are recommended to use only 2 attributes.