REPORT

"Tennis Match Winning Prediction"

Submitted by

Faiza Omar (14.02.04.044)

Group ID : 08
Lab Group No : A2

1 Introduction

Every year ATP arranges world class tennis tournaments in various cities for professional tennis players.By playing these tournaments, the players get huge prize money and ATP rankings, which provides a great impact on their carrier.Players always try to achieve trophies.In order to achieve trophies and ATP rankings, they always try to play as many matches as possible and get prepared for the grand slams tournaments.

2 Dataset

This dataset is build to predict whether a match is won by the player or not. A match is won because of some reasons and facts and of course great performances by the players.This dataset contains 2 classes, where result is 1 when a player wins the match and 0 if he looses the match. Each classes contains 104 instances. In here, winning is predicted of a player based on the performance of 2 ATP tournaments 2017. [1].

3 Models

I have used 5 models in this dataset to see the which one gives better accuracy & result.As K-Fold Cross Validation with k=5 has been applied, the dataset got divided into 5 slices and classifiers have been run on different slices exactly 5 times and lists of the results are also kept of each classifier.

3.1 Random Forest Classifier

Random forest classifier is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.IT is simple and easily parallelized. set [2].

3.2 Stochastic Gradient Descent Classifier

SGD classifier implements regularized linear models with stochastic gradient descent (SGD) learning: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate). SGD allows minibatch partial_fit method. For best results using the default learning rate schedule, the data should have zero mean and unit variance. [3].

3.3 Support Vector Classification

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall [4].

3.4 Naive Bayes Classifier

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes theorem with the naive assumption of independence between every pair of features.These have worked quite well in many real-world situations, famously document classification and spam filtering. They require a small amount of training data to estimate the necessary parameters. [5].

3.5 ExtraTrees Classifier

Extremely randomized trees (extra-trees) are another class of ensemble methods specifically designed for decision tree classifiers.Each tree is built from the original learning sample. At each test node, the best split is determined among random splits, and each one is determined by a random selection of an input (without replacement) and a threshold. [6].

4 Performance Scores of Models

4.1 Random Forest Classifier

Avg Accuracy	63%
Avg Precision	63%
Avg Recall	63%
Avg F1 Score	61%

4.2 SGD Classifier

Avg Accuracy	65%
Avg Precision	64%
Avg Recall	65%
Avg F1 Score	63%

4.3 Support Vector Classifier

Avg Accuracy	67%
Avg Precision	66%
Avg Recall	66%
Avg F1 Score	65%

4.4 Gaussian Naive Bayes Classifier

Avg Accuracy	68%
Avg Precision	69%
Avg Recall	70%
Avg F1 Score	67%

4.5 ExtraTrees Classifier

Avg Accuracy	66%
Avg Precision	64%
Avg Recall	65%
Avg F1 Score	64%

5 Discussion

From the above used models, Gaussian Naive Bayes Classier gives the best result.In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Ensemble learning helps improve machine learning results by combining several models.

References

[1] ATP World Tour. Atp world tour, 2017.

[2] Scikit Learn. Random forest classifier, 2018.

[3] Scikit Learn. Sgd classifier, 2018.

[4] Wikipedia. Support vector machine — wikipedia, the free encyclopedia, 2018.

[5] Scikit Learn. Gaussiannb classifier, 2018.

[6] Scikit Learn. Extratrees classifier, 2018.