Classification of unlabeled LoL match records with “Win/Loss”

1.Introduction:

League of Legends (LoL) is one of the most played eSports in the world at the moment. It is an online , 5 vs. 5 competitive PC game. In this game, there are many different fields that may change the game outcomes including the creation time, game duration and the number of towers, inhibitor each team has.

Though it has the uncertainty of the game outcomes, we can try to investigate what are the better strategy to win this game. The project has access to about 3 million match records of solo gamers as training set. Each record contains all publicly available game statistics of a match played by some gamer. Then predict the result of the other 2 million such records.

In this project, some classifiers are used to get the result of the prediction, including Decision Tree, Support Vector Machine and K Nearest Neighbor.

2.Algorithm:

2.1 Decision Tree(DT)

Decision tree is a method of machine learning. Decision tree is a kind of tree structure, in which each internal node represents a judgment on an attribute, each branch represents the output of a judgment result, and finally each leaf node represents a classification result. It is a case-based inductive learning.

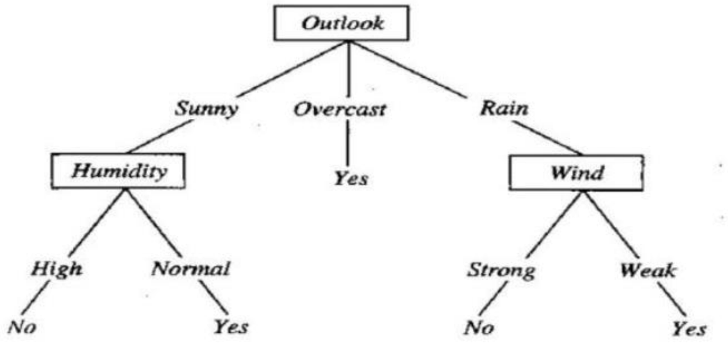


Figure 2.1.1 A simple example of Decision Tree structure

The parameters of the decision tree are shown in the table below(Other parameters which are not shown in the table is determined by default) :

Parameter	value
splitter	random
max_depth	15

Splitter is feature classification criteria, we can use the value ‘best’ for less data and use the value ‘random’ for large data.

Max_depth is the maximum depth of the decision tree, when the data is large, we should limit the value between 10 and 100.

2.2 K Nearest Neighbor(KNN)

KNN is a machine learning algorithm which can be used for classification and regression. For a given test sample, the nearest K training samples in the training set are found based on the distance measurement, and then the prediction is made based on the information of the K "neighbors".

In the classification task, voting method can be used to select the most frequent category markers in the K samples as the prediction result; in the regression task, the average value of the real value output markers of the K samples can be used as the prediction result. In addition, we can also use weighted average method based on distance.

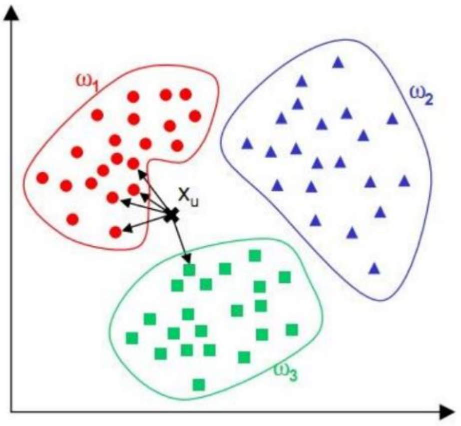


Figure2.2.1 An example of KNN

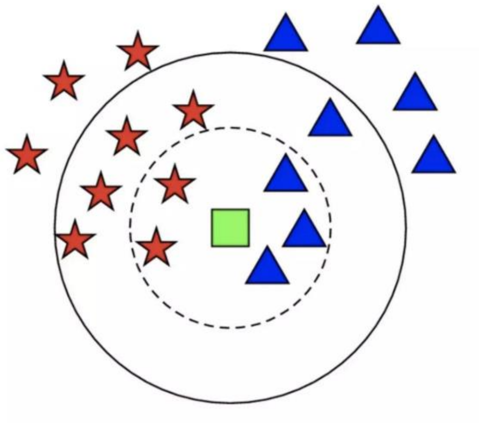


Figure2.2.2 An example of KNN

The parameters of the K Nearest Neighbor are shown in the table below(Other parameters which are not shown in the table is determined by default) :

Parameter	value

n_neighbors	5
weights	uniform
algorithm	auto

N_neighbors is the k value in the KNN.

Weights is used to identify the weight of the nearest neighbor samples of each sample, the default value is "uniform", and all the nearest neighbor samples have the same weight.

Algorithm is the algorithm of the nearest neighbor method with limited radius, the default value is ‘auto’.

2.3 Support Vector Machine(SVM)

Support vector machines (SVM) is a binary classification model. Its basic model is the linear classifier with the largest interval defined in the feature space. The maximum interval makes it different from the perceptron; SVM also includes kernel techniques, which makes it a non-linear classifier in essence. The learning strategy of SVM is to maximize the interval, which can be formalized as a convex quadratic programming problem, which is equivalent to the minimization of the regularized hinge loss function. The learning algorithm of SVM is the optimization algorithm to solve convex quadratic programming.

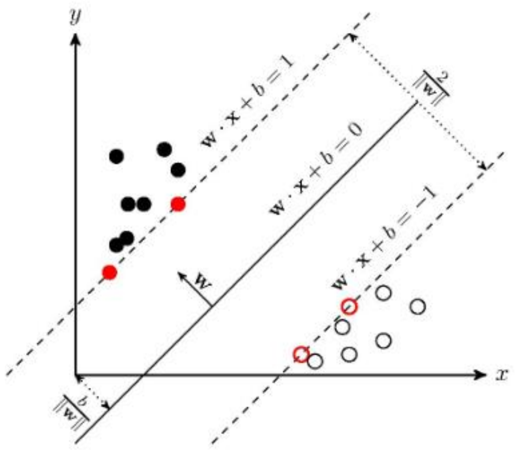


Figure2.3.1 An example of SVM

The parameters of the SVM are shown in the table below(Other parameters which are not shown in the table is determined by default) :

Parameter	**value**
kernel	rbf

gamma	auto

Kernel is the type of sum function used in the algorithm; the default value is rbf.

Gamma is the coefficient of kernel function, the default value is ‘auto’, the larger the gamma, the smaller the σ, which makes the Gaussian distribution high and thin, resulting in the model can only act near the support vector, which may lead to over fitting; on the contrary, the smaller the gamma, the larger the σ, the smoother the Gaussian distribution, and the poor classification effect on the training set may lead to under fitting.

3.Requirement

3.1 sklearn

Sklearn is a common module in machine learning. It encapsulates common machine learning methods, including regression, dimensional reduction, classification, clustering and so on.

3.2 numpy

Numpy provides Python support for multidimensional array objects. It supports a large number of advanced dimensional array and matrix operations, and also provides a large number of mathematical function libraries for array operations.

3.3 time

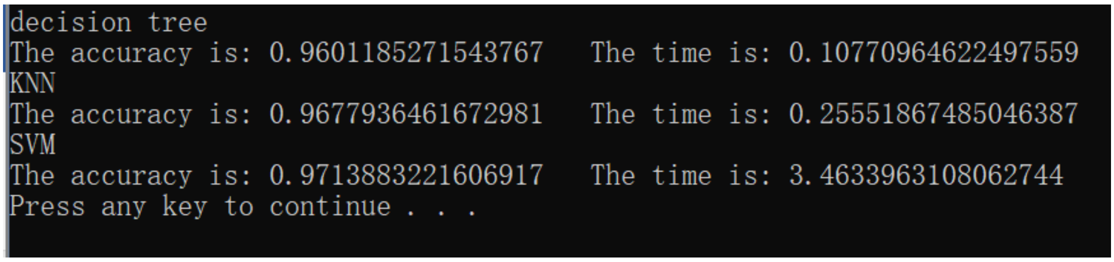
The time library is the standard library for processing time in Python. The time library can express the computer time, provide the method to obtain the system time and format the output, and provide the system level precise timing function (which can be used for program performance analysis).

3.4 pandas

Panels: panel data analysis. Pandas is built based on numpy, which provides a good support for time series analysis. There are two main data structures in pandas, one is Series and the other is DataFrame.

4.Results

4.1 the result from the screen



```
decision tree
The accuracy is: 0.9601185271543767   The time is: 0.10770964622497559
KNN
The accuracy is: 0.9677936461672981   The time is: 0.25551867485046387
SVM
The accuracy is: 0.9713883221606917   The time is: 3.4633963108062744
Press any key to continue . . .
```

4.2 summarize the result

	accuracy	Training time(s)
Decision Tree	0.9601185271543767	0.10770964622497559
K Nearest Neighbor	0.9677936461672981	0.25551867485046387
Support Vector Machine	0.9713883221606917	3.4633963108062744

5,Comparison and discussion

5.1 Comparison

From the result, we can tell that the Decision Tree and K Nearest Neighbor have less training time than Support Vector Machine. However, the accuracy of each classifier is nearly the same. In addition, different parameters can get different accuracies and training time.

5.2 Discussion

For the Decision Tree, we should pay attention to the parameter max_depth, when the data is large, we should limit the value between 10 and 100 to avoid overfitting.

For the K Nearest Neighbor, we should pay attention to the parameter n_neighbors, that is the K value. The smaller K value means that the whole model becomes complex and prone to over fitting. The larger K value means that the overall model becomes simple and prone to under fitting.