MATH4995 Report

Project 1-Titanic: Machine Learning on predicting survival of Titanic's passengers by using SVM model

Name: Lee Cheuk Yin

Introduction

Given the passengers' information of Titanic. This project requires me to predict the survival of each of them in file "test.csv" based on their corresponding information. In order to accomplish this task, the technique in machine learning is applied on this project. This report will show the process of using machine learning and analysis of the performance of the SVM model.

Data Preview

The datasets consist of 2 files, "train.csv" and "test.csv". In order to build the SVM model for accomplishing this task, we need to use the dataset in train.csv for training the SVM model. The train.csv consist of 891 passengers and their corresponding information that having 11 features and a label that indicating whether the specific passenger is alive or not.

Data Preprocessing

1. Counting numbers of unfilled value in the features

After counting, there are 687 unfilled value in feature "Cabin", 177 unfilled value in feature "Age" and 2 unfilled value in feature "Embarked".

2. Handling unfilled value found in above

There are only 2 unfilled value in feature "Embarked", which means I can just completely delete those 2 passengers. Compare to the total number of passengers provided in "train.csv" which is 891. It will not affect much on dataset.

For the "Age" feature, there are so many unfilled values. Just simplify deletes these passengers' information will seriously harm the accuracy of the model. Therefore, instead of simply deleting them, filling these unfilled values with the median of the remaining passenger's ages will be better approach.

The way of dealing feature "Cabin" will be showed in the following part of the report.

3. Digitize the Features that are belonged to Categorical Variable

There are some features that have no numerical value but categorical values, such as "Sex" and "Embarked". "Sex" feature has 'male' and 'female' as its values and

“Embarked” feature has ‘C’, ‘S’, ‘Q’ as its values. I can use 1 and 0 to represent female and male respectively. Similarly, I can use 0, 1 and 2 to represent ‘C’, ‘S’ and ‘Q’ respectively. The data will look like the following after the above actions are completed.

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	0	22.0	1	0	A/5 21171	7.2500	NaN	2
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	1	38.0	1	0	PC 17599	71.2833	C85	1
2	3	1	3	Heikkinen, Miss. Laina	1	26.0	0	0	STON/O2. 3101282	7.9250	NaN	2
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	1	35.0	1	0	113803	53.1000	C123	2
4	5	0	3	Allen, Mr. William Henry	0	35.0	0	0	373450	8.0500	NaN	2
5	6	0	3	Moran, Mr. James	0	28.0	0	0	330877	8.4583	NaN	0
6	7	0	1	McCarthy, Mr. Timothy J	0	54.0	0	0	17463	51.8625	E46	2
7	8	0	3	Palsson, Master. Gosta Leonard	0	2.0	3	1	349909	21.0750	NaN	2
8	9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	1	27.0	0	2	347742	11.1333	NaN	2
9	10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	1	14.0	1	0	237736	30.0708	NaN	1
10	11	1	3	Sandstrom, Miss. Marguerite Rut	1	4.0	1	1	PP 9549	16.7000	G6	2
11	12	1	1	Bonnell, Miss. Elizabeth	1	58.0	0	0	113783	26.5500	C103	2
12	13	0	3	Saundercock, Mr. William Henry	0	20.0	0	0	A/5. 2151	8.0500	NaN	2
13	14	0	3	Andersson, Mr. Anders Johan	0	39.0	1	5	347082	31.2750	NaN	2
14	15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	1	14.0	0	0	350406	7.8542	NaN	2
15	16	1	2	Hewlett, Mrs. (Mary D Kingcome)	1	55.0	0	0	248706	16.0000	NaN	2
16	17	0	3	Rice, Master. Eugene	0	2.0	4	1	382652	29.1250	NaN	0
17	18	1	2	Williams, Mr. Charles Eugene	0	28.0	0	0	244373	13.0000	NaN	2
18	19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vande...	1	31.0	1	0	345763	18.0000	NaN	2
19	20	1	3	Masselmani, Mrs. Fatima	1	28.0	0	0	2649	7.2250	NaN	1
20	21	0	2	Fynney, Mr. Joseph J	0	35.0	0	0	239865	26.0000	NaN	2
21	22	1	2	Beesley, Mr. Lawrence	0	34.0	0	0	248698	13.0000	D56	2
22	23	1	3	McGowan, Miss. Anna "Annie"	1	15.0	0	0	330923	8.0292	NaN	0
23	24	1	1	Sloper, Mr. William Thompson	0	28.0	0	0	113788	35.5000	A6	2
24	25	0	3	Palsson, Miss. Torborg Danira	1	8.0	3	1	349909	21.0750	NaN	2
25	26	1	3	Asplund, Mrs. Carl Oscar (Selma Augusta Emilia...	1	38.0	1	5	347077	31.3875	NaN	2
26	27	0	3	Emir, Mr. Farred Chehab	0	28.0	0	0	2631	7.2250	NaN	1
27	28	0	1	Fortune, Mr. Charles Alexander	0	19.0	3	2	19950	263.0000	C23 C25 C27	2
28	29	1	3	O'Dwyer, Miss. Ellen "Nellie"	1	28.0	0	0	330959	7.8792	NaN	0
29	30	0	3	Todoroff, Mr. Lalio	0	28.0	0	0	349216	7.8958	NaN	2

Data table (partial)

4. Show Data Summary

At this point, I can show the data summary such as correlation matrix and data statistics.

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
PassengerId	1	-0.00502832	-0.0353298	-0.043136	0.0313186	-0.0576859	-0.00165658	0.0127032	0.0305551
Survived	-0.00502832	1	-0.335549	0.541585	-0.0698217	-0.03404	0.0831508	0.25529	-0.108669
Pclass	-0.0353298	-0.335549	1	-0.127741	-0.336512	0.0816556	0.0168245	-0.548193	-0.0438347
Sex	-0.043136	0.541585	-0.127741	1	-0.0865058	0.116348	0.247508	0.179958	-0.118593
Age	0.0313186	-0.0698217	-0.336512	-0.0865058	1	-0.232543	-0.171485	0.0937071	0.00716521
SibSp	-0.0576859	-0.03404	0.0816556	0.116348	-0.232543	1	0.414542	0.160887	0.0606061
Parch	-0.00165658	0.0831508	0.0168245	0.247508	-0.171485	0.414542	1	0.217532	0.0793198
Fare	0.0127032	0.25529	-0.548193	0.179958	0.0937071	0.160887	0.217532	1	-0.0634623
Embarked	0.0305551	-0.108669	-0.0438347	-0.118593	0.00716521	0.0606061	0.0793198	-0.0634623	1

Correlation matrix

	PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Embarked
count	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000	889.000000
mean	446.000000	0.382452	2.311586	0.350956	29.315152	0.524184	0.382452	32.096681	1.637795
std	256.998173	0.486260	0.834700	0.477538	12.984932	1.103705	0.806761	49.697504	0.636157
min	1.000000	0.000000	1.000000	0.000000	0.420000	0.000000	0.000000	0.000000	0.000000
25%	224.000000	0.000000	2.000000	0.000000	22.000000	0.000000	0.000000	7.895800	1.000000
50%	446.000000	0.000000	3.000000	0.000000	28.000000	0.000000	0.000000	14.454200	2.000000
75%	668.000000	1.000000	3.000000	1.000000	35.000000	1.000000	0.000000	31.000000	2.000000
max	891.000000	1.000000	3.000000	1.000000	80.000000	8.000000	6.000000	512.329200	2.000000

Data Summary

Except the “Cabin”, “Name” and “Ticket”, the other features with digital values in it have been plotted in the correlation matrix and the data summary.

5. Data Cleansing

From common sense and the data table shown in above, I can see that the “PassengerId”, “Ticket” and “Name” can be removed before feeding the dataset into SVM model, because the name, the tickets that they are holding and the “PassengeId” of passengers has no relationship with their survival. From step 3 above, the feature “Cabin” has 687 unfilled values. Also, from the data table shown above, it is very hard to determine the value to be filled in those blanked space. Moreover, it doesn’t clearly show any obvious patterns and correlation with the survival label. Therefore, I can remove the feature “Cabin”.

One thing that is very important. Even feature “Embarked” has shown a huge correlation with survival. But including that in the model will generate unexpected weighting effect. Since the digitization of the categorical variable in “Embarked” provide a distinct value among 3 categories, the label of it will make a huge impact on the model. Due to the lack of numeric meaning to the model, the feature “Embarked” will also be removed. The following will be the resulting dataset.

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare
0	0	3	0	22.0	1	0	7.2500
1	1	1	1	38.0	1	0	71.2833
2	1	3	1	26.0	0	0	7.9250
3	1	1	1	35.0	1	0	53.1000
4	0	3	0	35.0	0	0	8.0500
5	0	3	0	28.0	0	0	8.4583
6	0	1	0	54.0	0	0	51.8625
7	0	3	0	2.0	3	1	21.0750
8	1	3	1	27.0	0	2	11.1333
9	1	2	1	14.0	1	0	30.0708
10	1	3	1	4.0	1	1	16.7000
11	1	1	1	58.0	0	0	26.5500
12	0	3	0	20.0	0	0	8.0500
13	0	3	0	39.0	1	5	31.2750
14	0	3	1	14.0	0	0	7.8542
15	1	2	1	55.0	0	0	16.0000
16	0	3	0	2.0	4	1	29.1250
17	1	2	0	28.0	0	0	13.0000
18	0	3	1	31.0	1	0	18.0000
19	1	3	1	28.0	0	0	7.2250
20	0	2	0	35.0	0	0	26.0000
21	1	2	0	34.0	0	0	13.0000
22	1	3	1	15.0	0	0	8.0292
23	1	1	0	28.0	0	0	35.5000
24	0	3	1	8.0	3	1	21.0750
25	1	3	1	38.0	1	5	31.3875
26	0	3	0	28.0	0	0	7.2250
27	0	1	0	19.0	3	2	263.0000
28	1	3	1	28.0	0	0	7.8792
29	0	3	0	28.0	0	0	7.8958

Resulting data table(partial)

6. Splitting the Dataset into test set and training set

Splitting the dataset in train set and data set into ratio = 0.7. (70% training data, 30% testing data)

7. Standardizing X_train and X_test

Finally, I need to standardize the X_train and X_test before fit them in the model. SVM model plays with kernel trick and try to project the original datapoints on even higher dimensions' hyperplane. If the data is too scattered, even the projection of datapoints on hyperplane will not be easy to find a solution that have the maximum margin. So, to solve this issue, the standardization of X_train and X_test is necessary. Otherwise, the SVM will not converge to a solution.

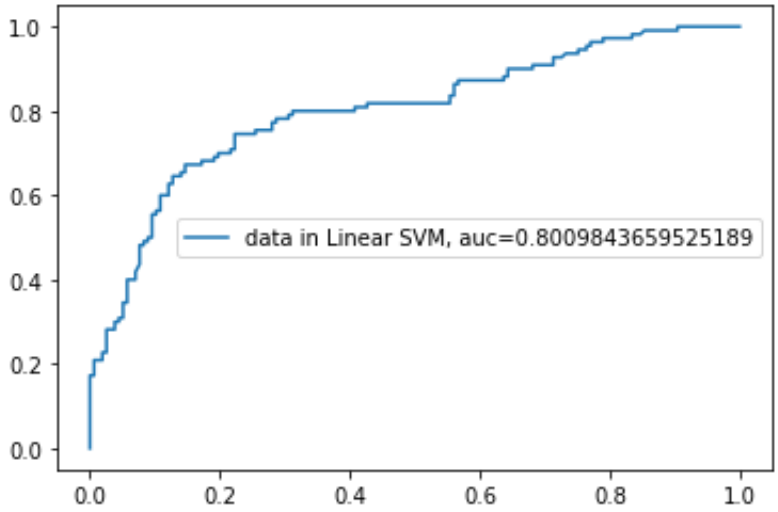
$$z = \frac{x_{ij} - \bar{x_i}}{\sigma}$$

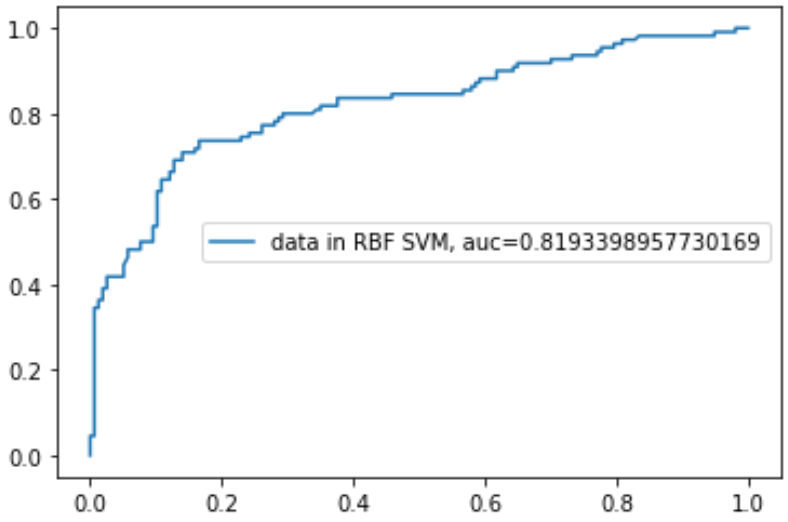
Formula of the standardization, x_{ij} is the j-th data in i-th feature, $\bar{x_i}$ is the mean of the data in i-th feature and σ is the standard deviation of the data in that feature.

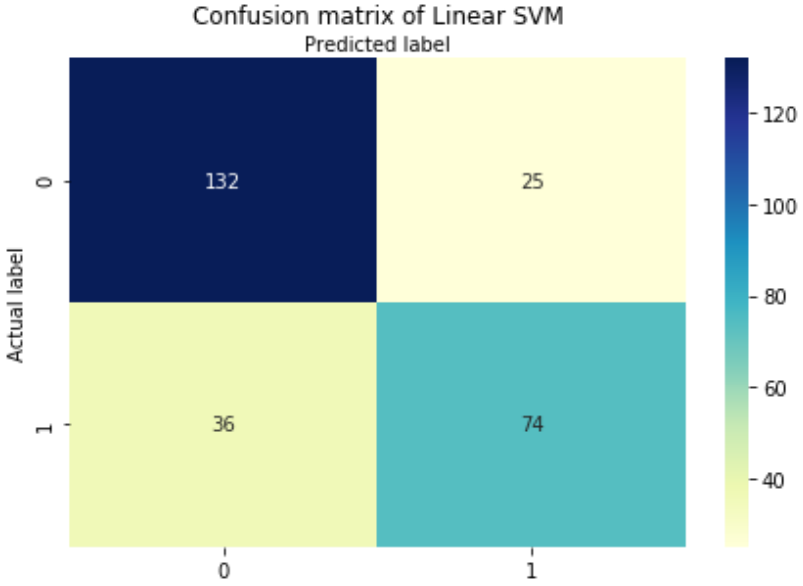
Data fitting using SVM

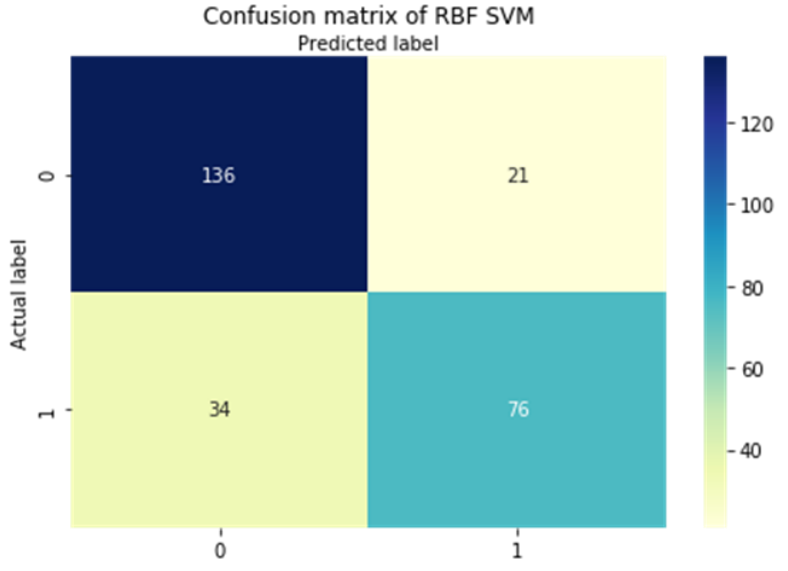
The Titanic survival problem is a typical binary classifying problem. It is because the label that is indicating the survival is either 0 or 1. SVM can use the kernel functions to project the data points on higher dimension's hyperplane which is fast in computation time and flexible in terms of different look of data because of variability of choice of kernel functions. In the following, the linear kernel and the radial basis function kernel will be used.

Model Performance




1.0
0.8
0.6
0.4
0.2
0.0
data in RBF SVM, auc=0.8193398957730169
0.0
0.2
0.4
0.6
0.8
1.0


Confusion matrix of Linear SVM
Predicted label
Actual label
132
25
36
74
0
1
120
100
80
60
40


Confusion matrix of RBF SVM
Predicted label
Actual label
136
21
34
76
0
1
120
100
80
60
40

```
Linear SVM Accuracy: 0.7715355805243446
Linear SVM Precision: 0.7474747474747475
Linear SVM Recall: 0.6727272727272727
RBF SVM Accuracy: 0.7940074906367042
RBF SVM Precision: 0.7835051546391752
RBF SVM Recall: 0.6909090909090909
```

Conclusion

From the above performance, using the linear or radial basis function kernel of SVM in Titanic Survival problem doesn't result the huge difference in terms of auc. But definitely, SVM with the radial basis function kernel performs better than linear one. Maybe the datapoints after the standardization have the geometry looks like a circle. Also, the recall score of both of them is significantly lower than other scores. This maybe the main reason of affecting the auc value. Next time, I will try to further reduce the number of features before using SVM on ML problem, because this action will favor the simplification of projection of SVM.