Machine Learning, Fall 2019: Project 3

Liam Li G48502460

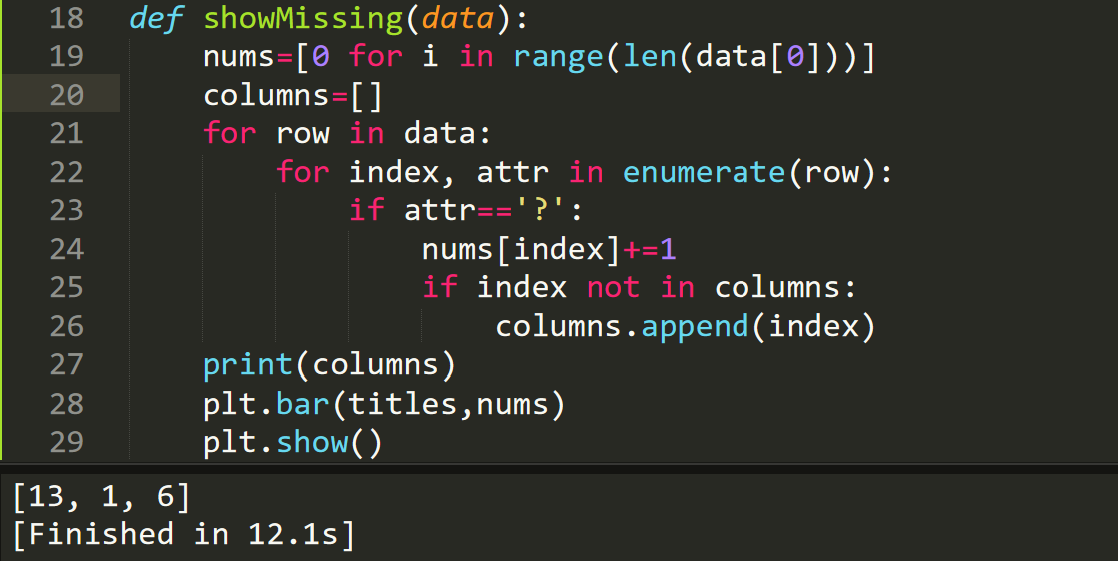
**Header:**

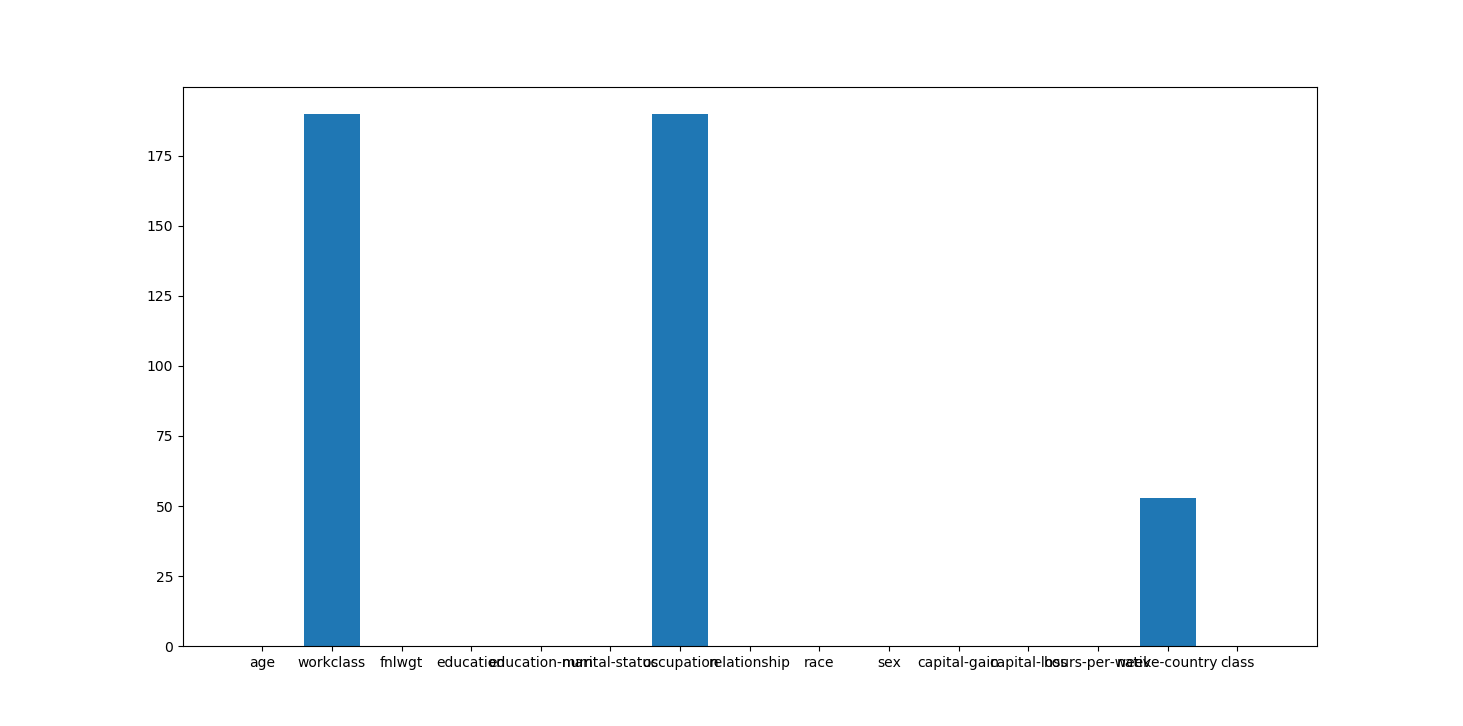
Python 3, sublime text 3

Modules: matplotlib, numpy, sklearn(SVC, linearSVC, RandomForest, learning\_curve, classification\_report), statistics(for modes), info\_gain, csv

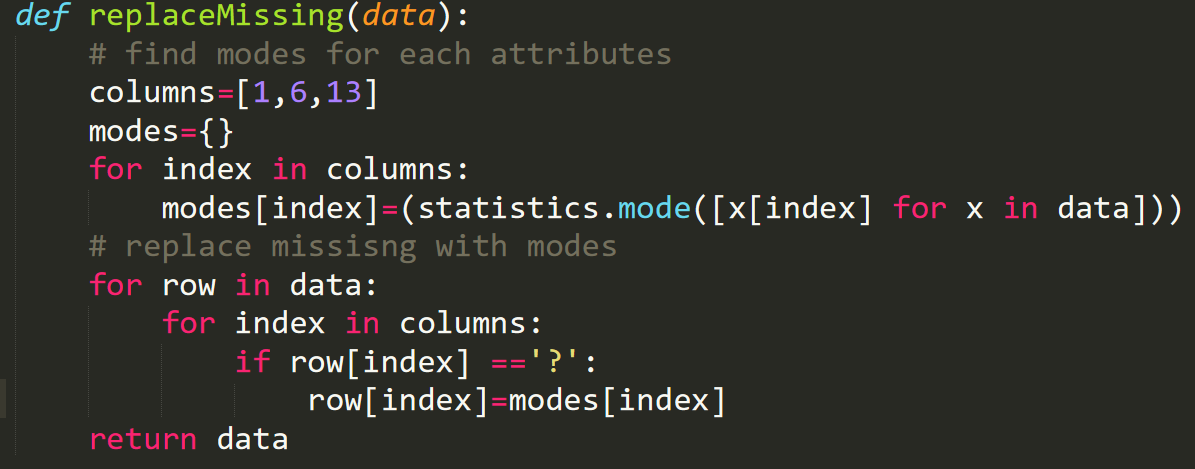
**Dataset preprocessing**

1. **Missing features**
2. Checking missing features





1. Replacing missing values with their modes.



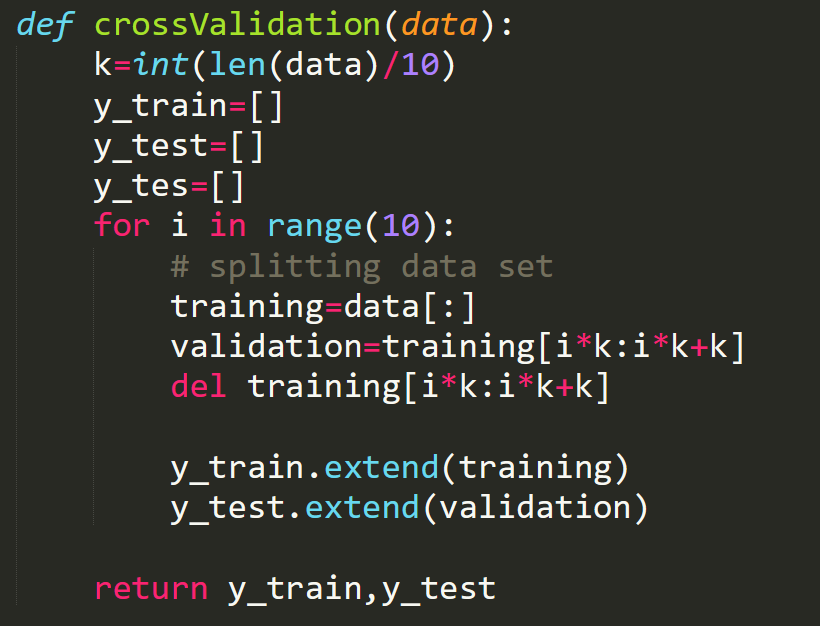
1. **Label Encoding**

It converts the nominal values into continuous number so the classifer can handle the data to train and test.

For some specific features like race and native country, united states and white are seen as 0 and other values are represented by 1 because of the small size of other values.

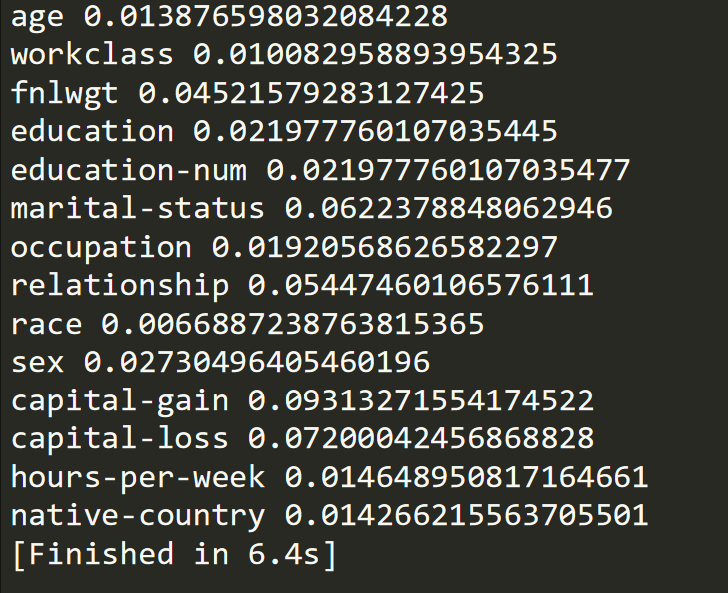
|  |  |
| --- | --- |
| Work class | {'State-gov': 0, 'Self-emp-not-inc': 1, 'Local-gov': 2, 'Self-emp-inc': 3, 'Private': 4, 'Federal-gov': 5, 'Without-pay': 6} |
| Education | {'Bachelors': 0, 'HS-grad': 1, 'Assoc-acdm': 2, 'Some-college': 3, '7th-8th': 4, 'Assoc-voc': 5, '5th-6th': 6, '11th': 7, '9th': 8, 'Masters': 9, '12th': 10, '10th': 11, 'Prof-school': 12, 'Doctorate': 13, '1st-4th': 14, 'Preschool': 15} |
| Marital status | {'Never-married': 0, 'Married-civ-spouse': 1, 'Divorced': 2, 'Widowed': 3, 'Separated': 4, 'Married-spouse-absent': 5, 'Married-AF-spouse': 6} |
| Occupation | {'Adm-clerical': 0, 'Farming-fishing': 1, 'Protective-serv': 2, 'Prof-specialty': 3, 'Exec-managerial': 4, 'Machine-op-inspct': 5, 'Tech-support': 6, 'Other-service': 7, 'Craft-repair': 8, 'Transport-moving': 9, 'Sales': 10, 'Handlers-cleaners': 11, 'Priv-house-serv': 12} |
| Relationship | {'Not-in-family': 0, 'Own-child': 1, 'Husband': 2, 'Unmarried': 3, 'Other-relative': 4, 'Wife': 5} |
| Race | {'White': 0, 'Black': 1, 'Other': 1, 'Asian-Pac-Islander': 1, 'Amer-Indian-Eskimo': 1} |
| Sex | {'Male': 0, 'Female': 1} |
| Country | {'United-States': 0,’Others’: 1} |
| Class | {'<=50K': 0, '>50K': 1} |

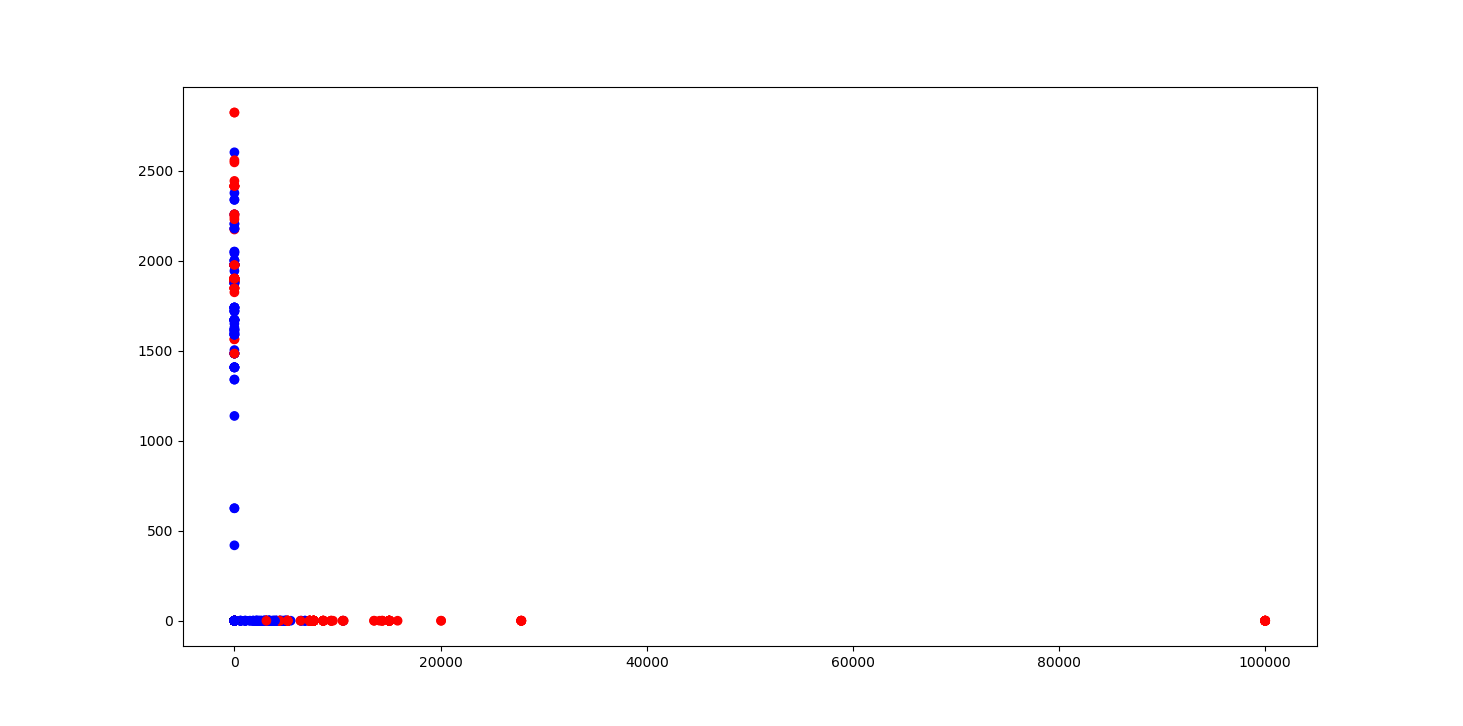
1. **10-fold-cross validation**



1. **Information gain**

The two highest information gain features are capital gain and capital loss.



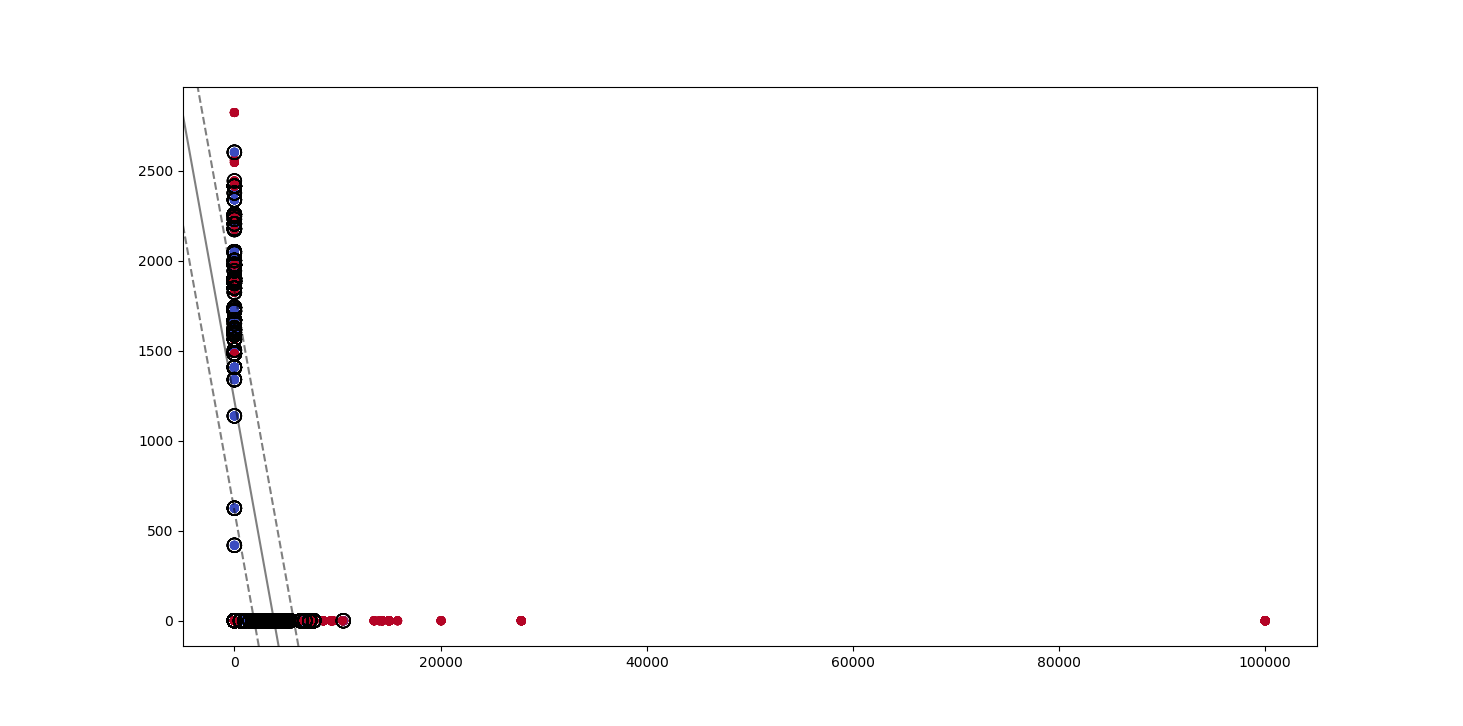


**Linear soft margin SVM**

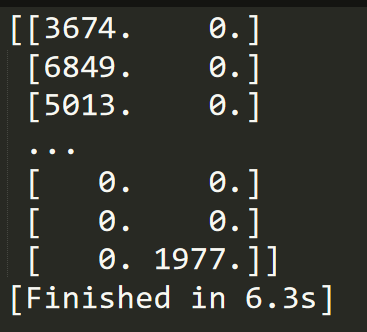
1. **Boundary visualization**

clf = SVC(C=1.0,gamma='scale',kernel=linear)

Support vectors are circled points.

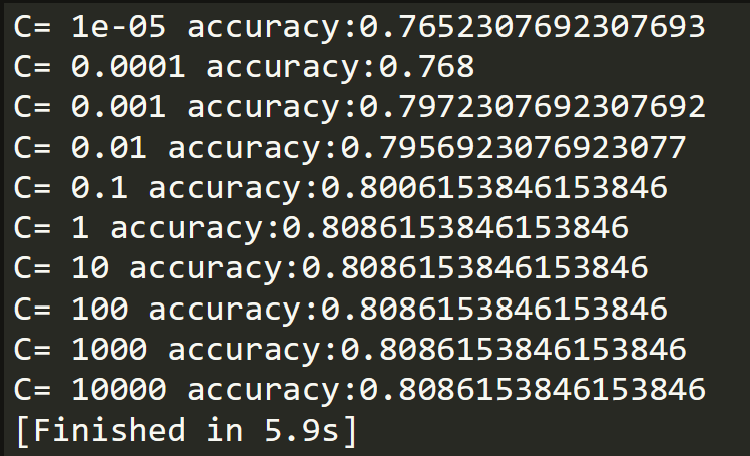


Some of the support vectors:

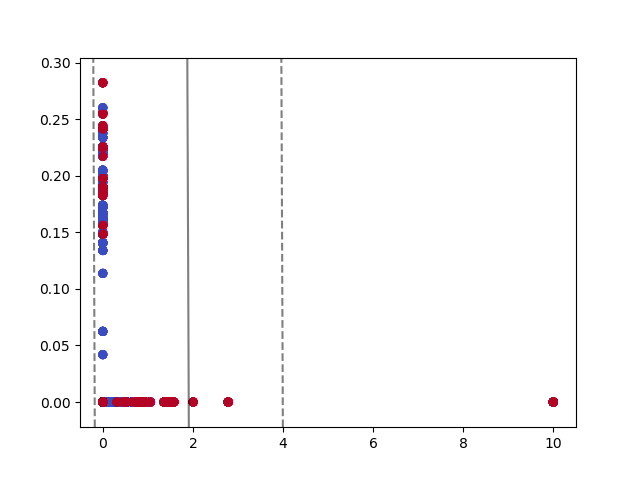


1. **Margins with different C**

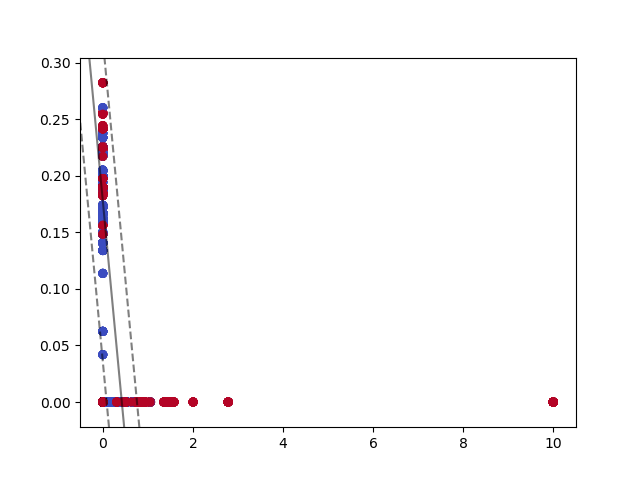
When I change the values of C from 10^-5 to 10^5 exponentially, the performance changes as follow. As we can see, the accuracy starts merging when C=0.1.



Decision boundary and margins when C=1e-4.



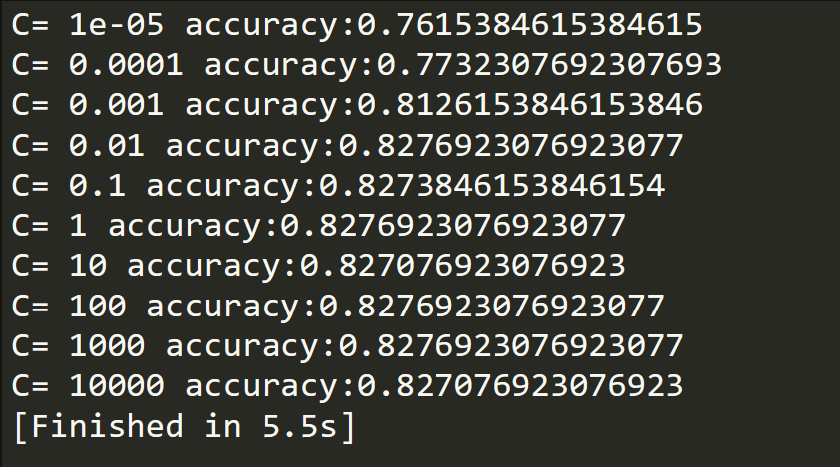
Decision boundary and margins when C=1.0.



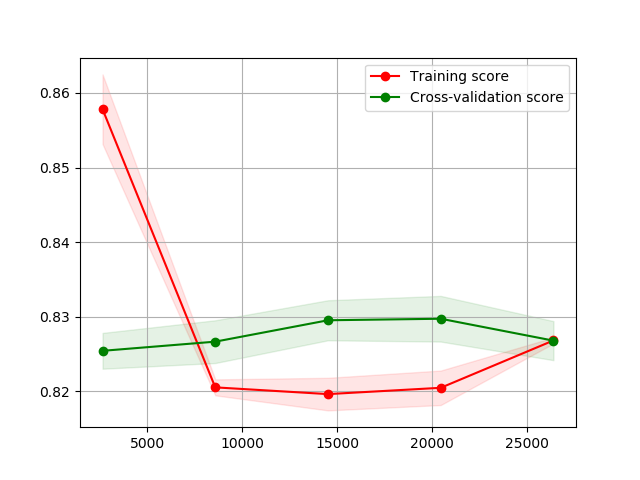
Effect: the smaller C brings larger margin and more misclassifications.

1. **SVM with all features**

Since we already know that C is convergence, I iteration different c values exponentially to get the optimal C for all features.



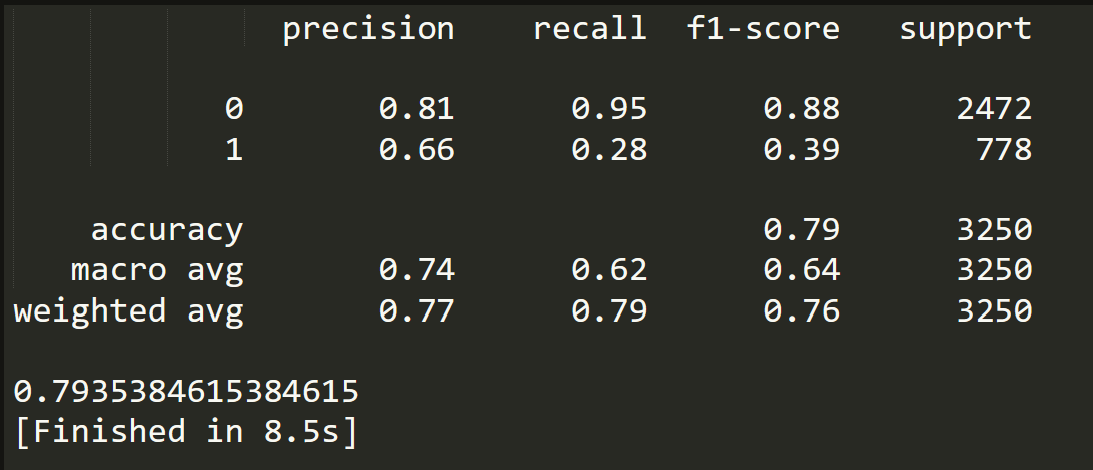
Learning curve when C=10000



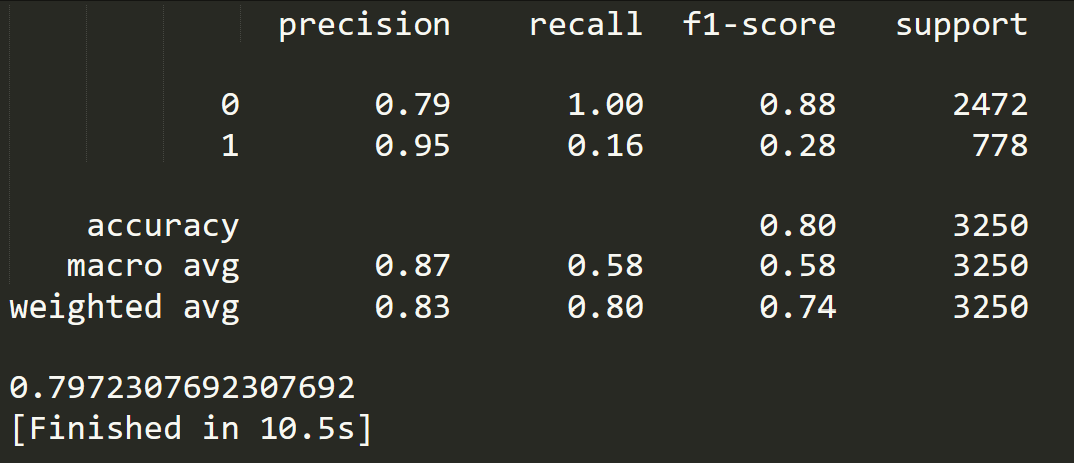
**SVM WITH KERNEL**

1. **Performance comparison**

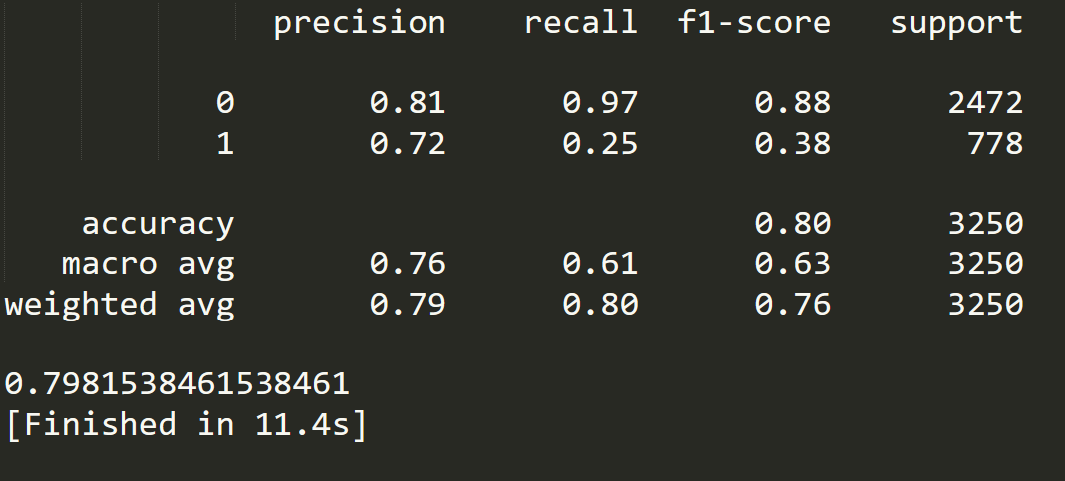
Linear: clf = SVC(C=1.0, gamma='scale',kernel='linear')



Polynomial: clf = SVC(C=1.0, gamma='scale',kernel='poly',degree=3)

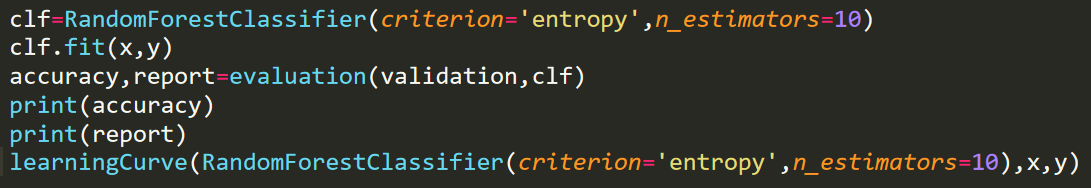


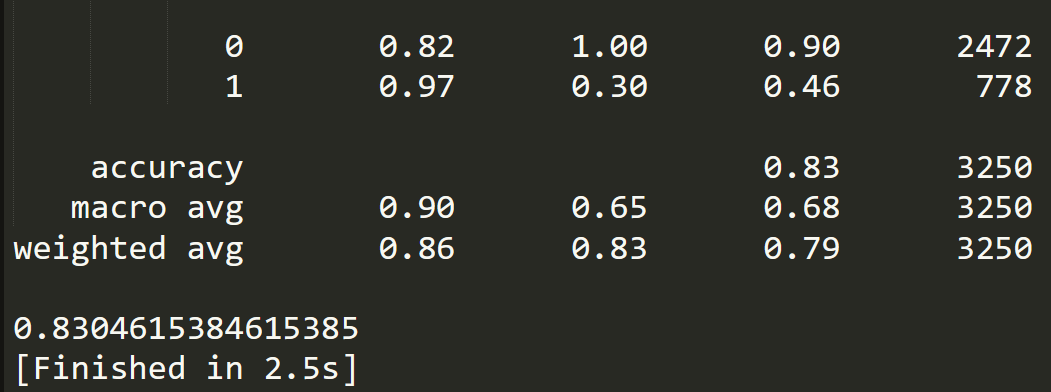
RBF: clf = SVC(C=1.0, gamma='auto',kernel='rbf')

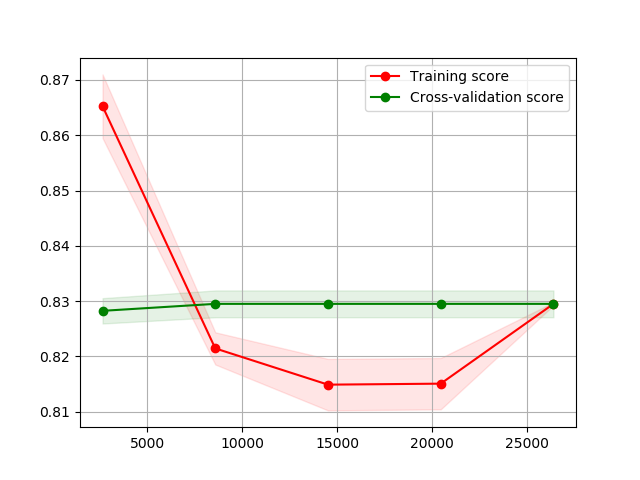


1. **Better performance**

Random forest: clf=RandomForestClassifier(criterion='entropy').







The cross validation curve doesn’t drop in the end compared with the linear SVM classfier, and the training score is slightly higher in the end.