Kaming
CS677 A1 Data Science with Python
Supervised Learning Summary
April 11, 2020

Supervised Learning Statistical Summary

No.	Classifier	Accuracy in Year 2	Profit in Year 2 (including the $100 principle)
1	SVM (d = 2)	73.58%	$100.00
2	Quadratic Discriminant	73.58%	$174.01
4	Linear Discriminant	75.47%	$137.62
3	SVM (linear)	75.47%	$182.09
6	Naïve Bayesian	79.25%	$171.04
5	Logistic Regression	79.25%	$183.91
7	SVM (gaussian)	81.13%	$182.30
8	Decision Tree	88.68%	$179.43
9	Random Forest	88.68%	$179.43
10	kNN	98.11%	$139.34

The statistical summary for the 10 supervised learning algorithms has been displayed in the table above. The chosen stock is YELP and it is the stock data in year 2018 as the training data and the task is to apply those different classifiers to classify the weekly labels in year 2019.

As indicated in the table, the accuracy for all 10 algorithms ranges from 73.58% to 98.11%, which is incredibly high for a supervised learning classifier. To be more specific, the most accurate classifier is k nearest neighbor (kNN) classifier, with 98.11% accuracy. Even the "worst" case, Polynomial SVM with degree = 2, has 73.58% accuracy. However, since I originally label the weekly data with specific and constant rules and the ratio of "green" to "red" label is around 3:7, which means that, even if any classifier predicts as all "red" labels (as Polynomial SVM with degree = 2 did), the accuracy will still be satisfied, I would not be surprised to see all those high accuracy results. But I am surprised when I see how accurate it is when applying kNN classifier to the data. One possible reason for this outstanding result is that all the data points are concentrated with the records in the same class. In other words, "green" labels are surrounded by other "green" labels, so are "red" labels. So, when kNN classifier counts k number of nearest records, it can get the correct result at the most of time.

Luckily, for the monetary results, it is always profitable in year 2019 if we invest $100 at the beginning of the year and trade with the predicted labels. The worst case is when we apply Polynomial SVM with degree = 2 and it turns out that all the labels are predicted as “red”, which forbids us to trade throughout the entire year and holds the $100 principle at the end of the year. However, the results also indicate that the most profitable method does not have to be the most accurate one. In this case, the most profitable method is to trade with the labels predicted by Logistic Regression model, with $183.91 at the end of the year but the accuracy against the “true label” is only 79.25%. On the other hand, the most accurate algorithm, kNN, only earns $139.34 with $100 investment at the beginning of the year.