ECE657A Assignment № 2

He Bing, 20848700 , b29he@uwaterloo.ca Feb 2020

Problem 1

Difference before and after normalization.

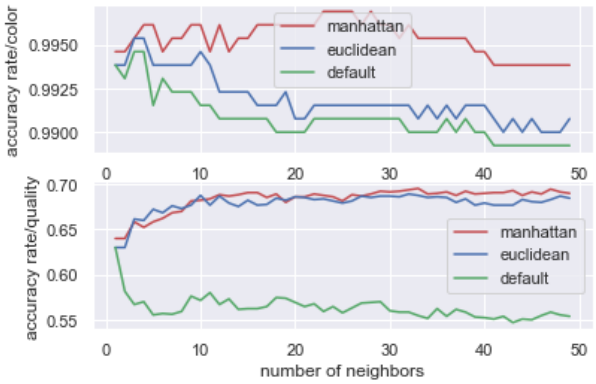
We use standard scaler function to normalize the data.Here are the two pairplots.
We can see that after normalization, the relative position of the points didn't change. Because all the thing that the transformer did is to standardize features by removing the mean and scaling to unit variance. In other words: $z = \frac{x-\mu}{s}$,(s is the standard deviation of training samples) so the relative position wouldn't change.

Figure 1: Pairplot before normalization.

Figure 2: Pairplot after normalization.

KNN classification performance under 3 kinds of models

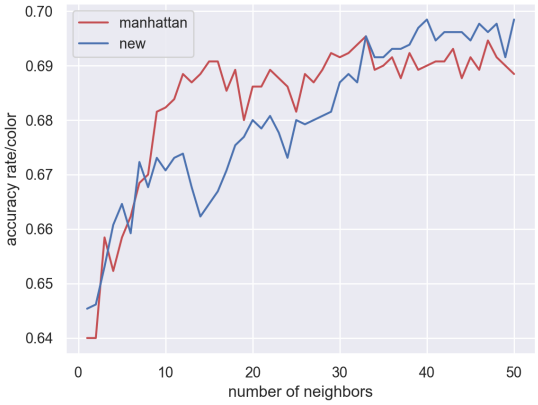
Figure 3: color / quality



Bonus 1: find additional schemes that better than manhattan+znorm

For this quesiton, I tried to improve the accuracy of classifying different qualities, for 'manhattan+znorm' does a great work on color classification. There is no optimization space for color. I tried many other preprocessing methods on the data. Then I found that manhattan+OrdinalEncoder works better than manhattan+znorm when number of neighbors between 33 and 50.

Figure 4: manhattan+znorm vs. manhattan+OrdinalEncoder



Bonus2: find a set of 4 that does better than all data on same metrics

I tried all the combinations within the 11 features using the following code.
Color and quality have different sets.

Listing 1: bonus 2

```
for s in itertools.combinations(D,4):
    X1_train, X1_test, y1_train,y1_test=train_test_split(StandardScaler().↩
       fit_transform(wine[list(s)].values), y1, test_size=0.2,↩
       random_state=ran)
    result1 = []
    count=0
    for n in neighbors:
        neigh_1=KNeighborsClassifier(n_neighbors=n)
        neigh_1.fit(X1_train,y1_train)
        result1.append(neigh_1.score(X1_test,y1_test))
    for r in range(len(all_accuracy)):
        if result1[r] > all_accuracy[r]:
            count+=1
    if count > 10:
        print(str(count),s)
```

color

Then I found that under the default condition, there is only one combination that is better than all the features, which is (’residual sugar’, ’chlorides’, ’total sulfur dioxide’, ’density’).

Figure 5: bonus 2 for color

quality

The combination is (’volatile acidity’, ’total sulfur dioxide’, ’density’, ’alcohol’). And the plot is as follow.

Figure 6: bonus 2 for quality

Compare the performance between PCA and LDA

(a) Performance under default (b) Performance under manhattan (c) Performance under euclidean

Figure 7: PCA vs LDA

LDA works better than PCA under default configuration, but is not so good as that in the other two situations. What's more, LDA is more stable than PCA. But the accuracy of PCA will decrease as the number of neighbors grows.

k plots of different feature sets

Figure 8: All features. color/quality

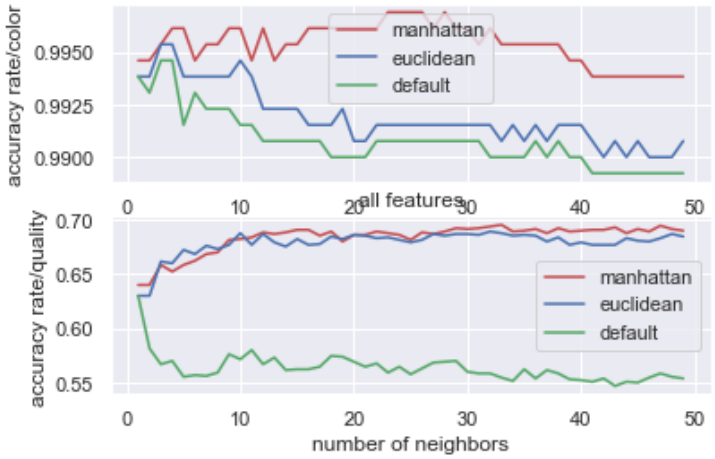


Figure 9: First 5 features. color/quality

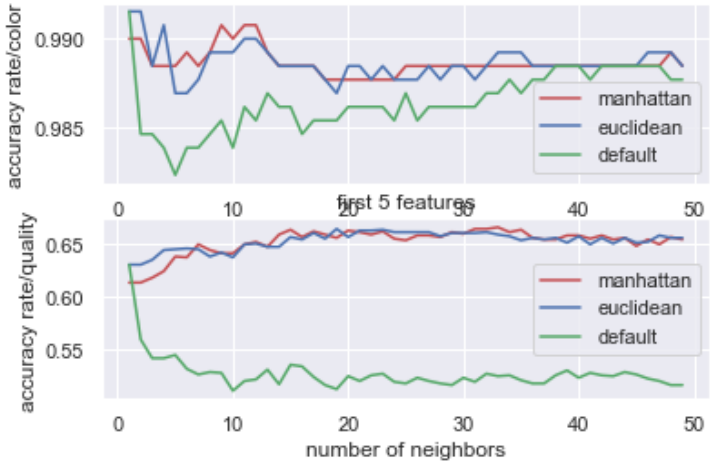
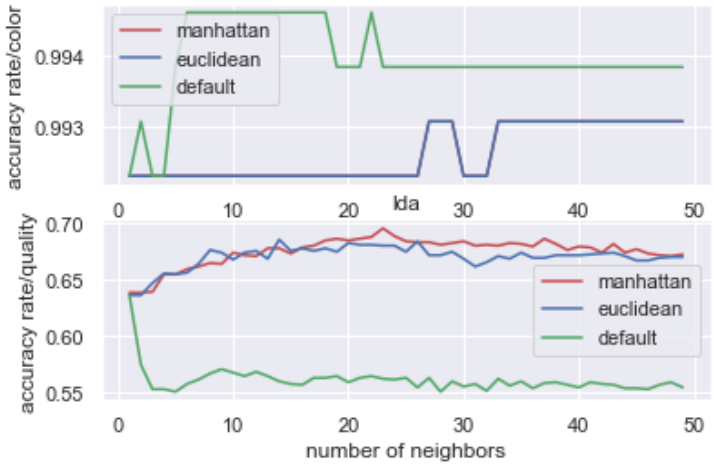


Figure 10: LDA. color/quality



Some discussion on the relationship between the features from any analysis you performed.

The set of all features under manhattan works better than any other feature set. In the mean time, the performance of first 5 features is very close to that of all features. And manhattan and euclidean weights functions are better than the default configuration in all features and first 5 features. But in LDA features, uniform is a little better than euclidean.

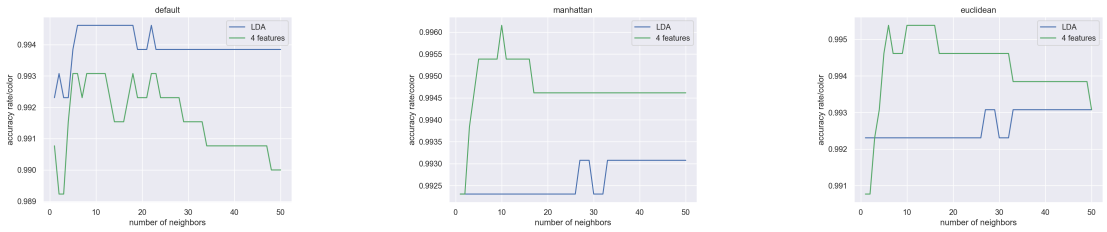
Selected Features

Were you able to find a subset of features that worked better than all features?

For color: ('residual sugar', 'chlorides', 'total sulfur dioxide', 'density')
For quality:('volatile acidity', 'total sulfur dioxide', 'density', 'alcohol')

How about compared to PCA or LDA?

As the graph shows:



(a) Performance under default (b) Performance under manhattan (c) Performance under euclidean

Figure 11: my set vs LDA

The feature set is not so good as LDA under default configuration. But it's better than LDA under manhattan and euclidean.

PCA vs. LDA

(The plots can be found in the previous part.)

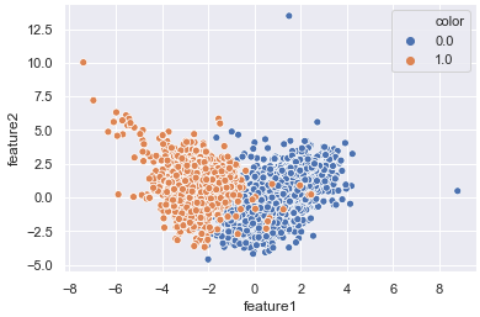
Did either of these methods help in this situation?Which worked better for the task?

From the figure, we can see PCA is as good as all features under default and euclidean, but LDA helps with the accuracy. But under manhattan, neither of PCA nor LDA help with this situation. So we can say LDA works better than PCA.

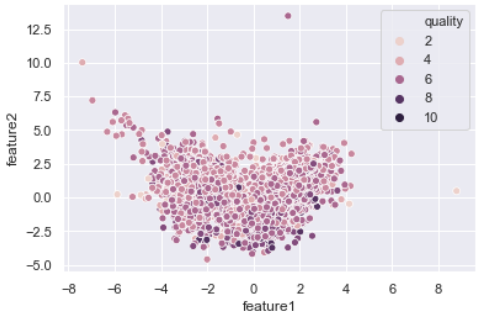
Did normalization impact the performance of either of them?

Yes. Normalization is an indispensable part of PCA. And it will improve the performance a lot. But for LDA, standardized and non-standardized data are exactly the same.

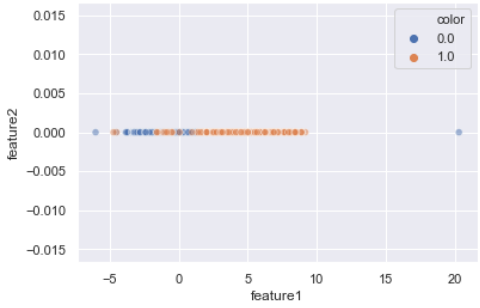
Plot



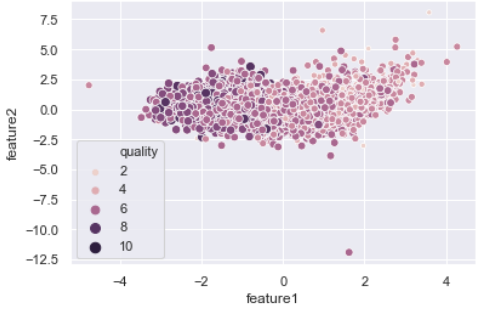
(a) PCA / color



(b) PCA / quality



(c) LDA / color



(d) LDA / quality

Figure 12: 4 plots

PCA can clearly split the points into 2 categories in color. And the classifying is better than most of the pair plots. So we know than PCA helps the points map into a better coordinators. As for LDA, because there are only 2 kinds of color, we can get a 1D scatter plot. LDA help the distance of means of the two clusters becomes longer. As for quality, LDA does a better job than PCA.