Predicting Quality of Wine

With k-Nearest-Neighbor Algorithm

B659 – APPLIED MACHINE LEARNING
PROJECT REPORT

December 9, 2014

Authored by: Ankit Sadana
(asadana@indiana.edu)

Predicting Quality of Wine

With k-Nearest-Neighbor Algorithm

Objective of this project

To use k-Nearest-Neighbor algorithm on two wine datasets (red and white), while altering the conditions of operation on this data to analyze the change in accuracy of prediction.

Datasets

Both datasets have been taken from the UCI database (link and citation at the end).

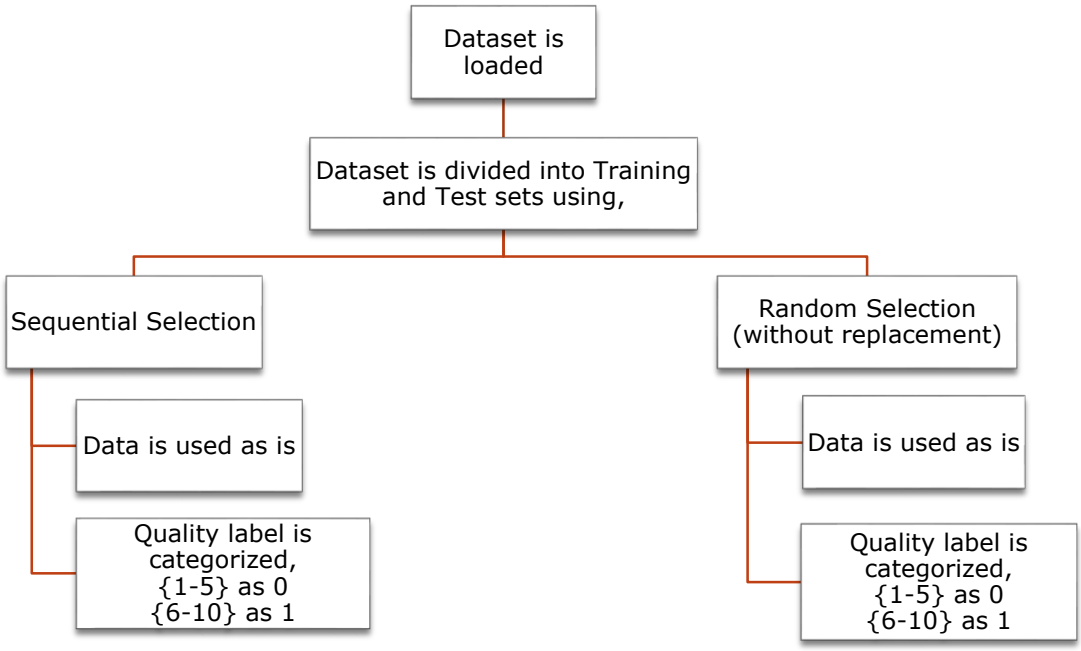
Red Wine Number of Instances: 1599 Number of Features: 11 + 1 Predicting Label		**White Wine** Number of Instances: 4898 Number of Features: 11 + 1 Predicting Label	
1. Fixed Acidity	continuous	1. Fixed Acidity	continuous
2. Volatile Acidity	continuous	2. Volatile Acidity	continuous
3. Citric Acid	continuous	3. Citric Acid	continuous
4. Residual Sugar	continuous	4. Residual Sugar	continuous
5. Chlorides	continuous	5. Chlorides	continuous
6. Free Sulfur Dioxide	continuous	6. Free Sulfur Dioxide	continuous
7. Total Sulfur Dioxide	continuous	7. Total Sulfur Dioxide	continuous
8. Density	continuous	8. Density	continuous
9. pH	continuous	9. pH	continuous
10. Sulphates	continuous	10. Sulphates	continuous
11. Alcohol	continuous	11. Alcohol	continuous
12. Quality	{1 to 10}	12. Quality	{1 to 10}

These datasets are not balanced, like there are much more middle range (5-6) quality examples as compared to the end of the range (1-3, 8-10) examples.

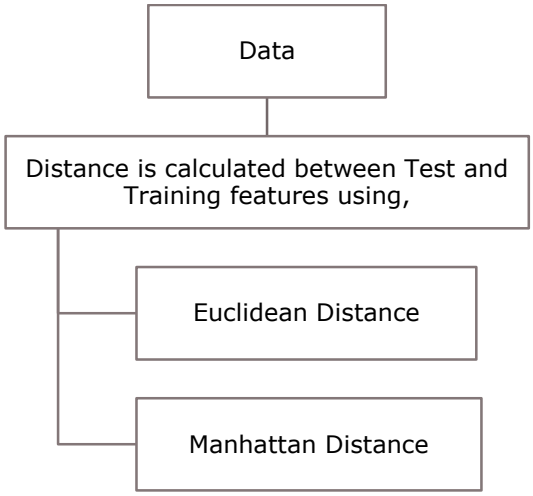
In the Red wine dataset, there is no instance of Quality values 1, 2, 9 and 10.

In the White wine dataset, there is no instance of Quality values 1, 2 and 10.

Methodology



- All these sets created above then undergo the following process,



- A matrix is created in each case with Test example number of columns and Training example number of rows. This matrix is then sorted in ascending order, column wise, giving us sorted distance of training examples for each test example.
- For each k = 1 to k = n, 'k' number of neighbors are selected for each Test example instance.
- A majority vote is taken between these neighbors to produce a prediction.
- This prediction is then compared to the Test label for each example in the Test set.
- Correct count is counted for each value of k and accuracy is returned by dividing it with the number of test examples.

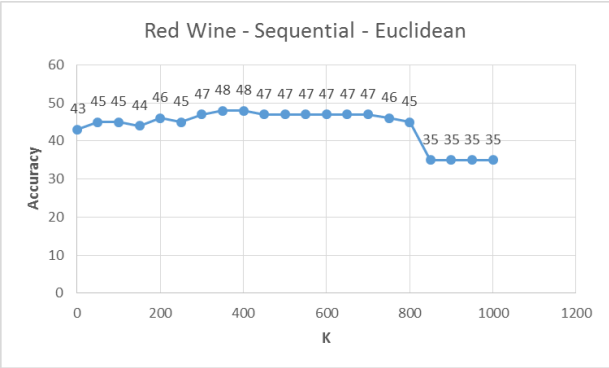
Results

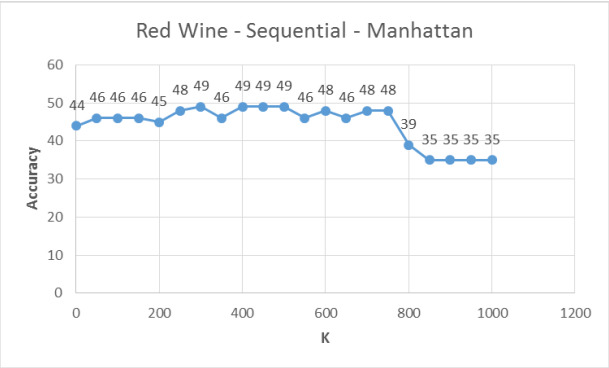
The results varied with variation of sampling methods, categorizing as well as changing the distance measure, but some patterns were also observed.

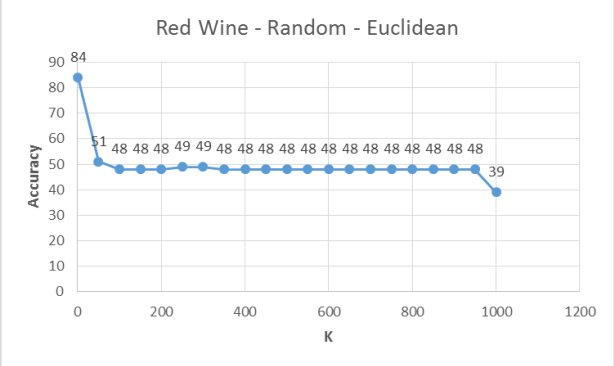
For convenience, all graphs start at K = 1 to K = Number of training examples (n), with K incrementing by 50.

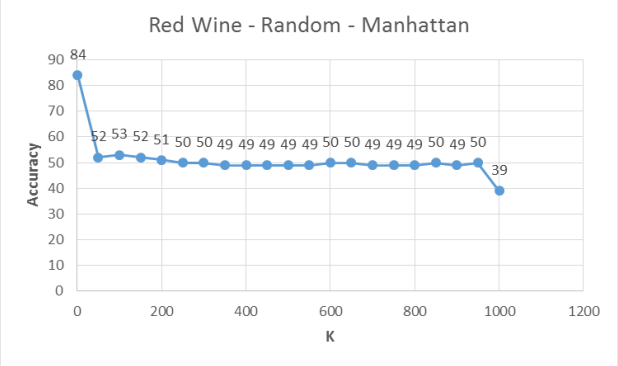
All graphs follow the Title naming: "Dataset – Sampling type – Distance measure".

Red Wine: Without Categorized Quality







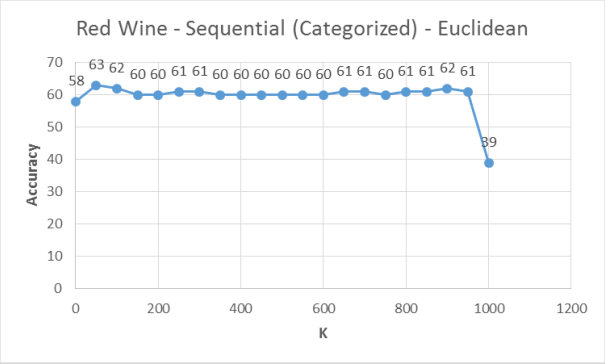


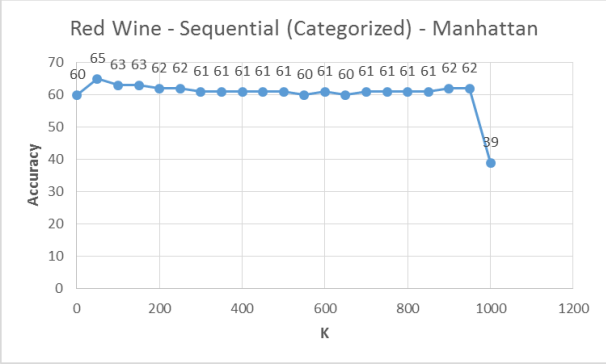
Observation:
Very small improvement is seen while using the Manhattan distance approach in these cases.

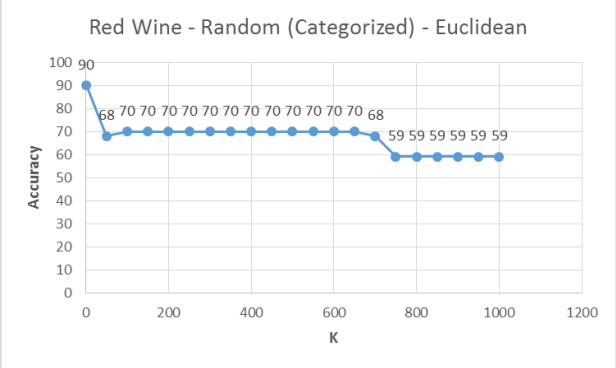
For the Sequential versus Random sampling approach, we observe that accuracy values have a distinct high at K = 1 and a distinct low for K = n.

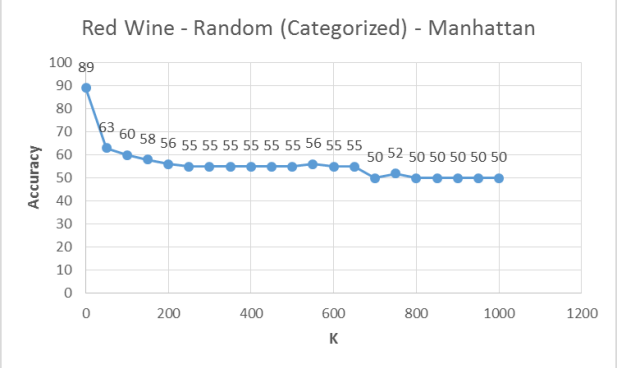
Both Euclidean and Manhattan seem to generate a much more linear graph for K vs Accuracy in case of Random Sampling.

Red Wine: With Quality Values Categorized







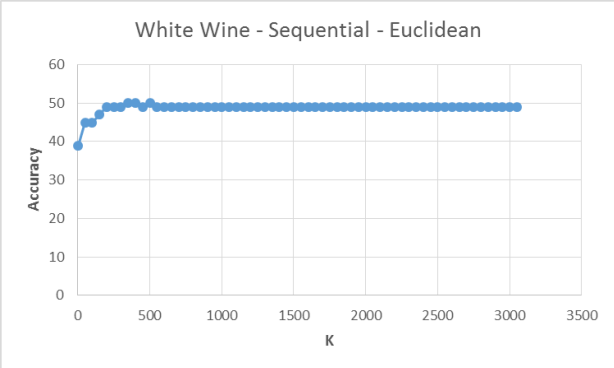


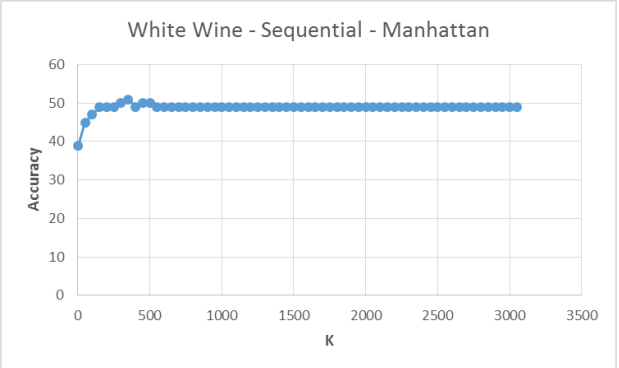
Observation:

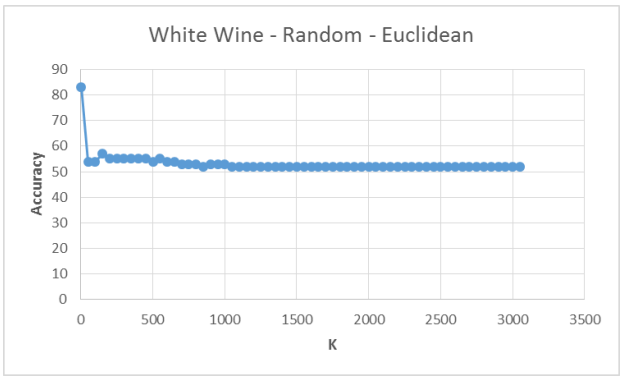
Manhattan distance measure performs mildly better for the Sequential sampling, but a Euclidean performs better in case of Random sampling.

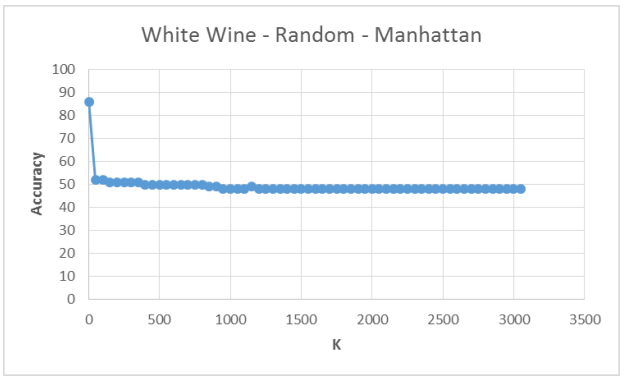
A distinct low is observed in Sequential sampling for K = n, which seems to disappear in Random sampling, but instead a distinct high in accuracy is observed for K=1 for Random sampling.

White Wine: Without Categorized Quality









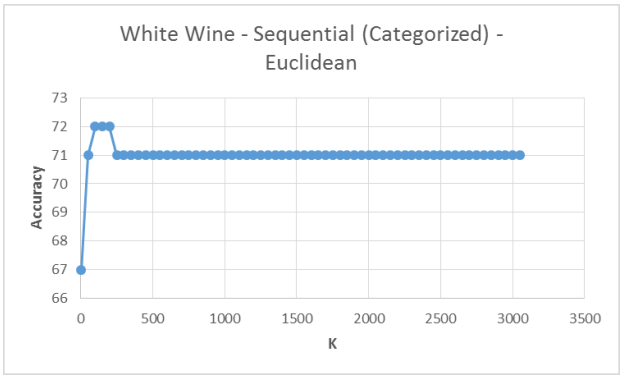
Observation:

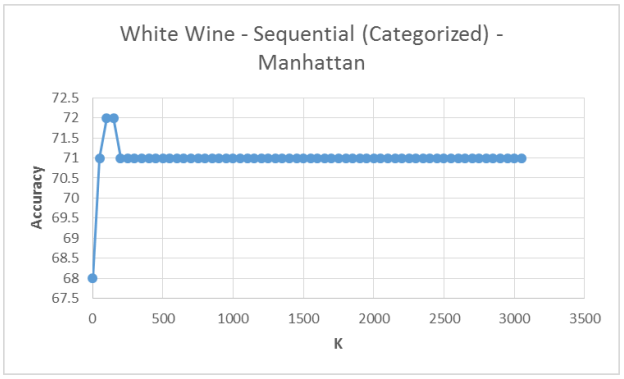
Both Euclidean and Manhattan distance measures perform nearly identically for both sequential and random sampling.

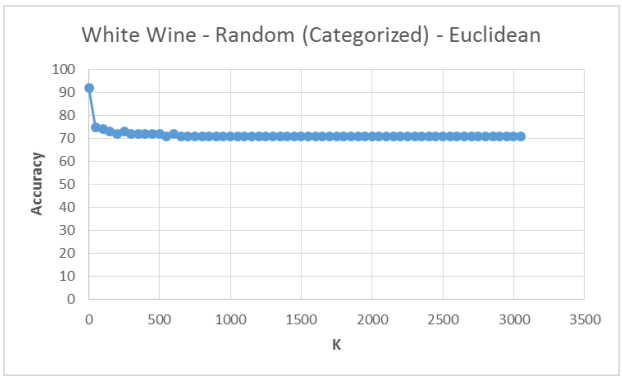
In sequential sampling, K=1 seems to be the lowest point of accuracy (under 40%) while for Random sampling, K=1 is the highest point of accuracy (closing to 90%).

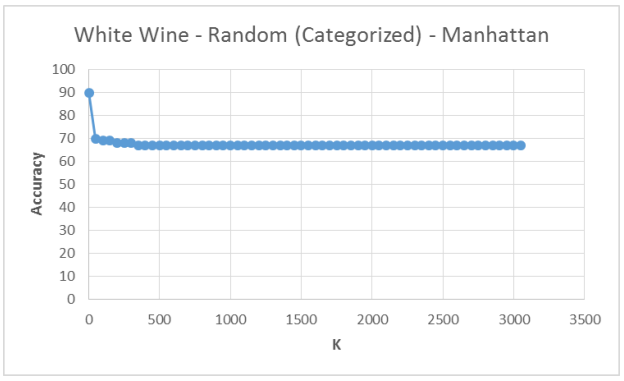
A very linear graph is observed in both cases, and accuracy remains almost the same while going from Sequential to Random.

White Wine: With Quality Values Categorized









Observation:
Both Euclidean and Manhattan distances perform nearly identically for both sequential and random sampling methods.

In sequential sampling, K=1 seems to be the lowest point of accuracy (~68%) while for Random sampling, K=1 is the highest point of accuracy (~90%).

A very linear graph is observed specially in Random after K=1, and accuracy remains almost the same while going from Sequential to Random.

Analysis

In this project we saw that both Euclidean and Manhattan distance metrics performed almost the same with Manhattan having a small advantage in most cases, this is because Euclidean distance metric uses square root approach to normalize the distances from all features. This even though necessary in some cases, seems to decrease the prediction accuracy in this case.

The Sequential sampling for both Red and White wine datasets gave us a low accuracy for K=1 and for Red wine a distinct low as 'K' approached 'n' (Number of training examples). The Random sampling for both Red and White wine, gave us a distinct high for K=1 and almost a linear graph towards 'K' approaching 'n'. This shows that Random sampling gives us a broader variety of examples, enabling the algorithm to find better neighbors for prediction, especially for 'K' closer to 1.

The linear nature of the graphs, for White wine, seems to be common in all graphs so it is safe to assume that this happens due to the nature of examples in the set. As for Red wine data, a general decrease in accuracy is observed in Sequential data, while in Random, it appears to be more gradual, but both appear around K = 700-800, showing that at this stage. This happens because at this point the majority class is predicted as it approaching 'n', making the graph appear linear.

Furthermore, categorizing the data increased the accuracy of prediction for k-Nearest Neighbor algorithm for about 10-20% for both set, Red and White wine. This was expected since the probability of prediction was (1/10) without categorization, and it became (1/2) with categorization.

Conclusion

When the data set is being split for training and testing, random sampling gives a more diverse variety of examples for training and testing, thereby improving the overall quality of the prediction. We also confirmed that Manhattan distance metric perform slightly better than the Euclidean distance metric for this data set, and as expected categorization of labels increases the prediction accuracy.

Predicting the Quality of Wine is an example of how use the ingredients of a product, a manufacturer can predict if the product will be well received, before the product undergoes manufacturing process. This might further be improved upon by applying a feature selection algorithm to test for irrelevant features and by testing other distance metrics' performance compared to Euclidean and Manhattan.

Sources

- Databases taken from: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

  P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

  Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

- Class Notes and Slides, given by Prof. Sriraam Natarajan
- Distance measures from: http://www.saedsayad.com/k_nearest_neighbors.htm
- Chapter 2: Nearest Neighbors, from "A course in Machine Learning" By: Hal Daumé III

End of Report