Theodore Nguyen Due July $10^{th}$ 11:55 pm

Introduction to Machine Learning
Assignment Paper #1

## Part I: Reflection on your learning (20pt)

(1) What have you learned in the class so far?

I've learned about the differences between supervised and unsupervised learning. I've also learned what `accuracy`, `precision`, and `recall` mean and how to calculate them. I've learned the general concept behind $k$-Nearest Neighbors and how to use Python to execute the algorithm.

(2) Which area (knowledge, skillset, attitude, etc.) of learning in this class do you think you are strong at and which area do you want to do better and how?

I think I am strong at quickly reading and comprehending Python code. I've taken an Introduction to Python class last summer, and the code in this class serves as a good reminder of its syntax. I want to dig deeper into my knowledge of the machine learning models and algorithms. I can do this by reading the course texts outside of class.

(3) Did you get help from your peers, or who did you collaborate with? Provide names and briefly describe their support/contributions to your assignment.

I got help from Megan Young. She has helped me to better understand cross validation and its purpose.

## Part II: KNN report (40pt) on a wine dataset (wine.csv)

(1) Briefly describe the data (i.e., the number of data points, variables, labels, etc.) 5pt

There are 14 columns (features) in this wine data set. The 14 features are:

(i) Wine

(ii) Alcohol

(iii) Malic.acid

(iv) Ash

(v) Acl

(vi) Mg

(vii) Phenols

(viii) Flavanoids

(ix) Nonflavanoid.phenols

(x) Pronanth

(xi) Color.int

(xii) Hue

(xiii) OD

(xiv) Proline

There are 178 data points (samples). The unique values in the Wine column are 1, 2, and 3. This means that every sample (data point) can be categorized into three categories: Wine 1, Wine 2, and Wine 3.

(2) Select three or four features for classification. Which ones did you select and why? 5pt

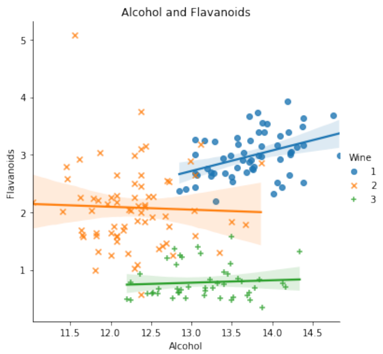
Chosen features: Alcohol, Flavanoids, Proline, Color.int

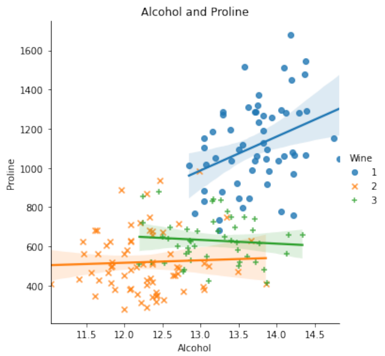
(i) Wines often vary in their alcohol content, so choosing the Alcohol feature could be useful.

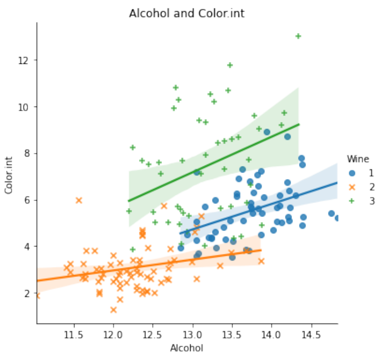
(ii) Flavanoids affect many aspects of wine including its color, taste, and texture, so Flavanoids can be another useful feature.

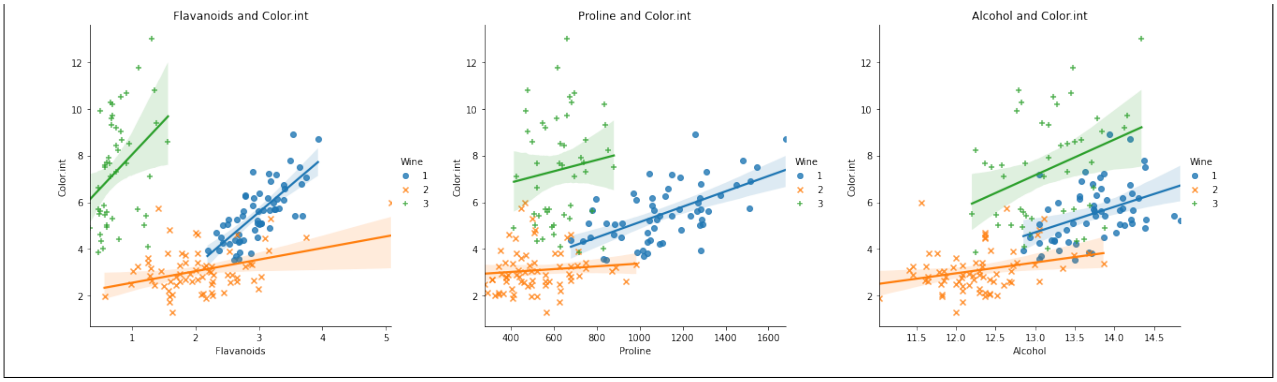
(iii) Proline is an amino acid that is prevalent in wine, so this feature can also be useful in helping us categorize various wines.

(iv) Different wines often have different and distinct colors, so Color.int can be a useful feature.









(3) How did you split (i.e., ratio) training vs. testing? 5pt

- Training: 80% (0.8)
  - Training size: 142 data points (samples)
- Testing: 20% (0.2)
  - Testing size: 36 data points (samples)

(4) Use "pickle" and produce two pkl files. Provide the file names here. 5pt

- `wines_train.pkl`
- `wines_test.pkl`

(5) How many $k$'s did you generate? 5pt

I generated 68 $k$'s.

(6) How many cross validation scores did you generate? 5pt

I generated 68 cross validation scores.

(7) What is the optimal value of $k$ and how did you decide? 5pt

The optimal value of $k$ is 26 because out of the 68 $k$'s (ranging from 3 to 70) it gave the highest cross validation score of 0.73429.

(8) How did your KNN perform? What's your accuracy score? 5pt

My KNN performed mediocrely with an accuracy of 0.72222.