HW1 – Red Wine Neural Network Classification

INF552 YUNJIE ZHAO

Part1 - Data

Number of Instances:	1599
Number of Attributes:	12

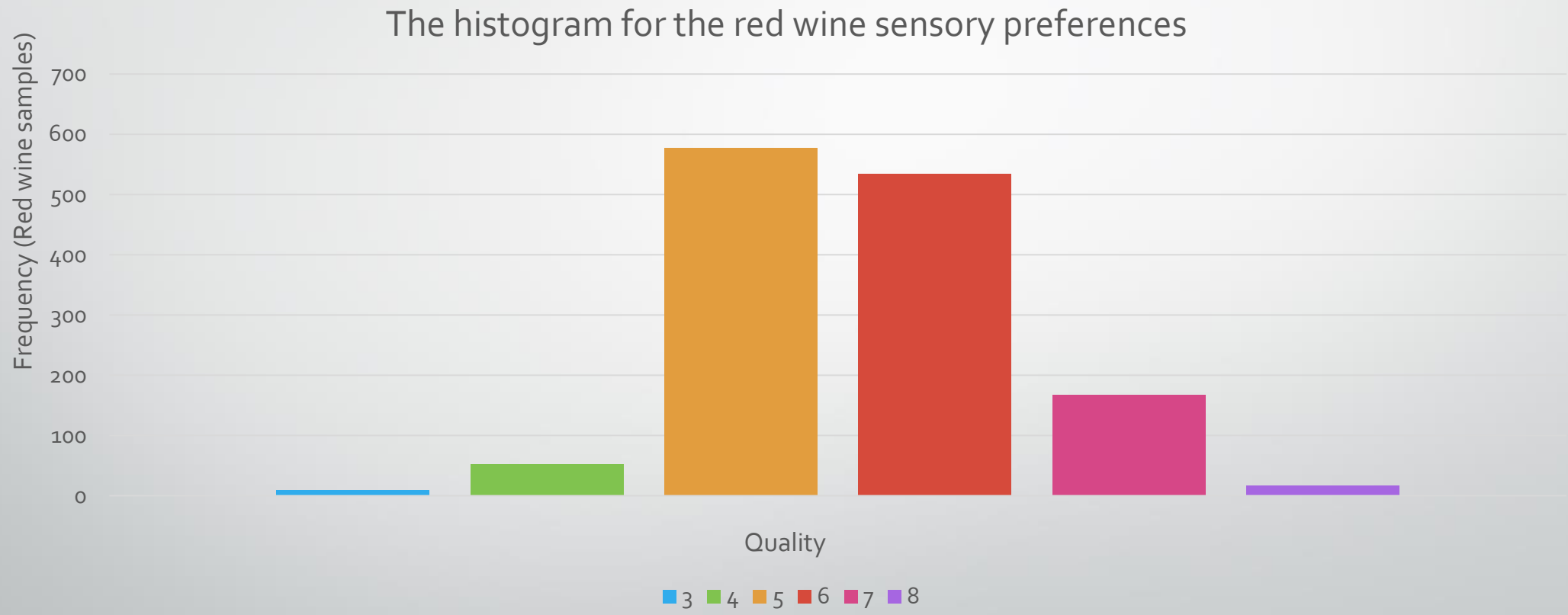
Attribute Information	
Input variables	Output variable
1 - fixed acidity	quality (score between 0 and 10)
2 - volatile acidity	
3 - citric acid	
4 - residual sugar	
5 - chlorides	
6 - free sulfur dioxide	
7 - total sulfur dioxide	
8 - density	
9 - pH	
10 - sulphates	
11 - alcohol	

Part1 - Data

Attribute (units)	Min	Max	Mean
fixed acidity (g(tartaric acid)=dm^3)	4.6	15.9	8.3
volatile acidity (g(acetic acid)=dm^3)	0.1	1.6	0.5
citric acid (g=dm^3)	0.0	1.0	0.3
residual sugar (g=dm^3)	0.9	15.5	2.5
chlorides (g(sodium chloride)=dm^3)	0.01	0.61	0.08
free sulfur dioxide (mg=dm^3)	1	72	14
total sulfur dioxide (mg=dm^3)	6	289	46
density (g=cm^3)	0.990	1.004	0.996
pH	2.7	4.0	3.3
sulphates (g(potassium sulphate)=dm^3)	0.3	2.0	0.7
alcohol (% vol.)	8.4	14.9	10.4

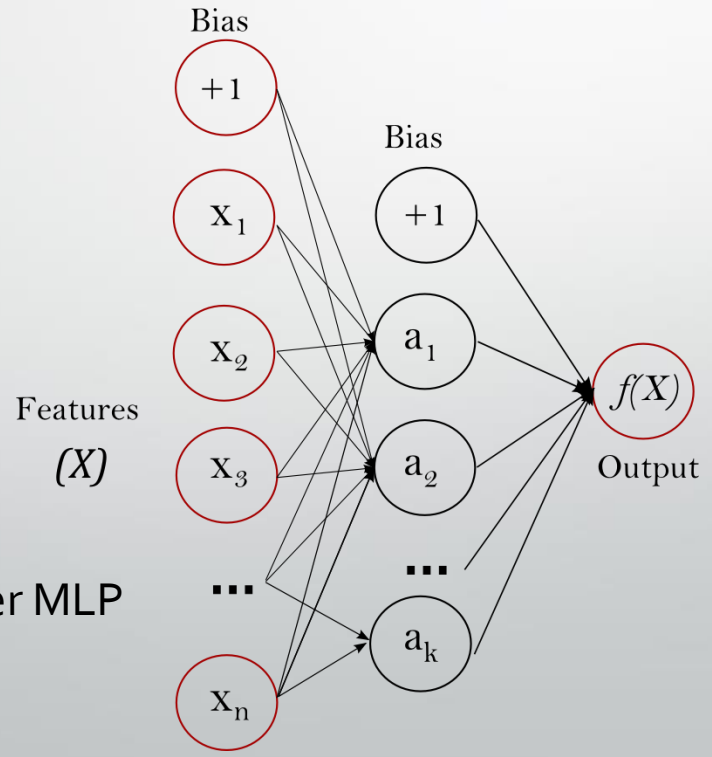
The Physicochemical Data Statistics

Part1 - Data


The histogram for the red wine sensory preferences
Frequency (Red wine samples)
700
600
500
400
300
200
100
0
Quality
3
4
5
6
7
8

Part2 - Algorithm

- Application: Scikit-Learn version 0.18
  - Classification: Neural Network - Multi-layer Perceptron
  - Given a set of features **X = {x_1, x_2, …, x_m}** and a target **y**, it can learn a non-linear function approximator for classification



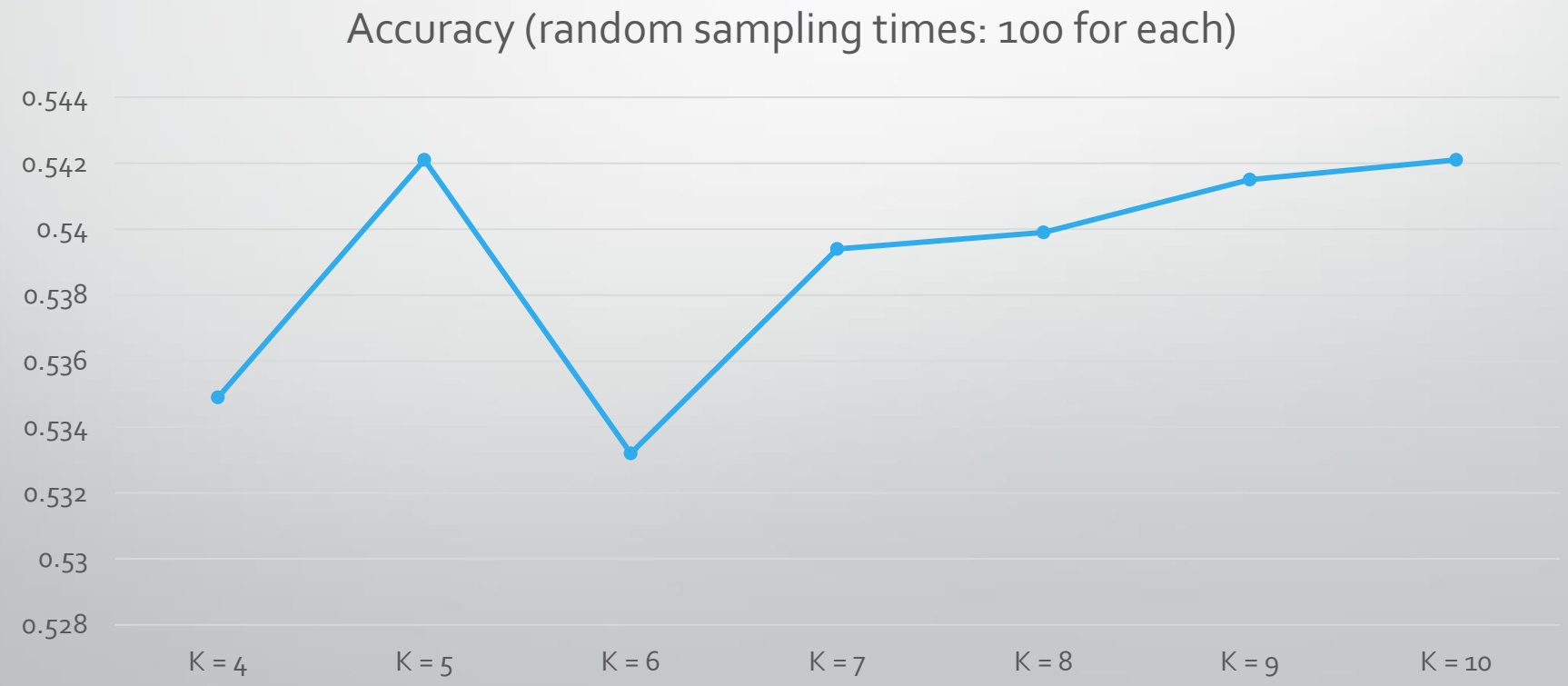
One hidden layer MLP

Part2 - Algorithm

- Step-1 Clean Data
  - Deduplication
    - Number of Instances: 1599 reduced to 1359
- Step-2 Training
  - MLPClassifier
    - a multi-layer perceptron (MLP) algorithm that trains using backpropagation.
    - Training Parameters: hidden layer sizes = 5, max iteration = 200
- Step-3 Prediction
  - Test Data
- Step-4 Evaluation
  - K-fold cross validation

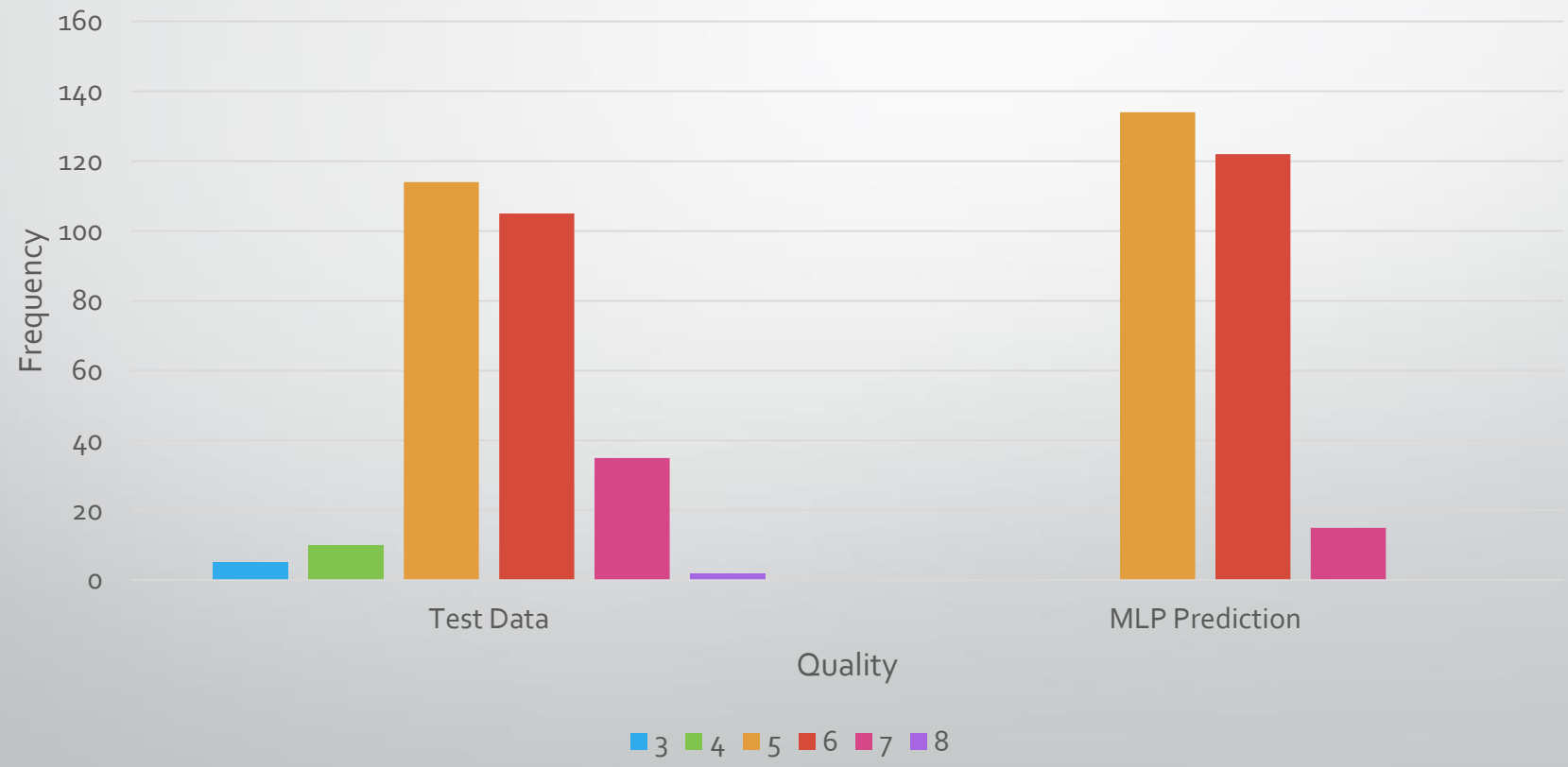
Part2 - Evaluation

- K-fold cross validation
  - Evaluate data with a different K(4 to 10)
  - In each K, do 100 random sampling. make a comparison between MLPClassication and Test data, and get a average score of 100 sampling.

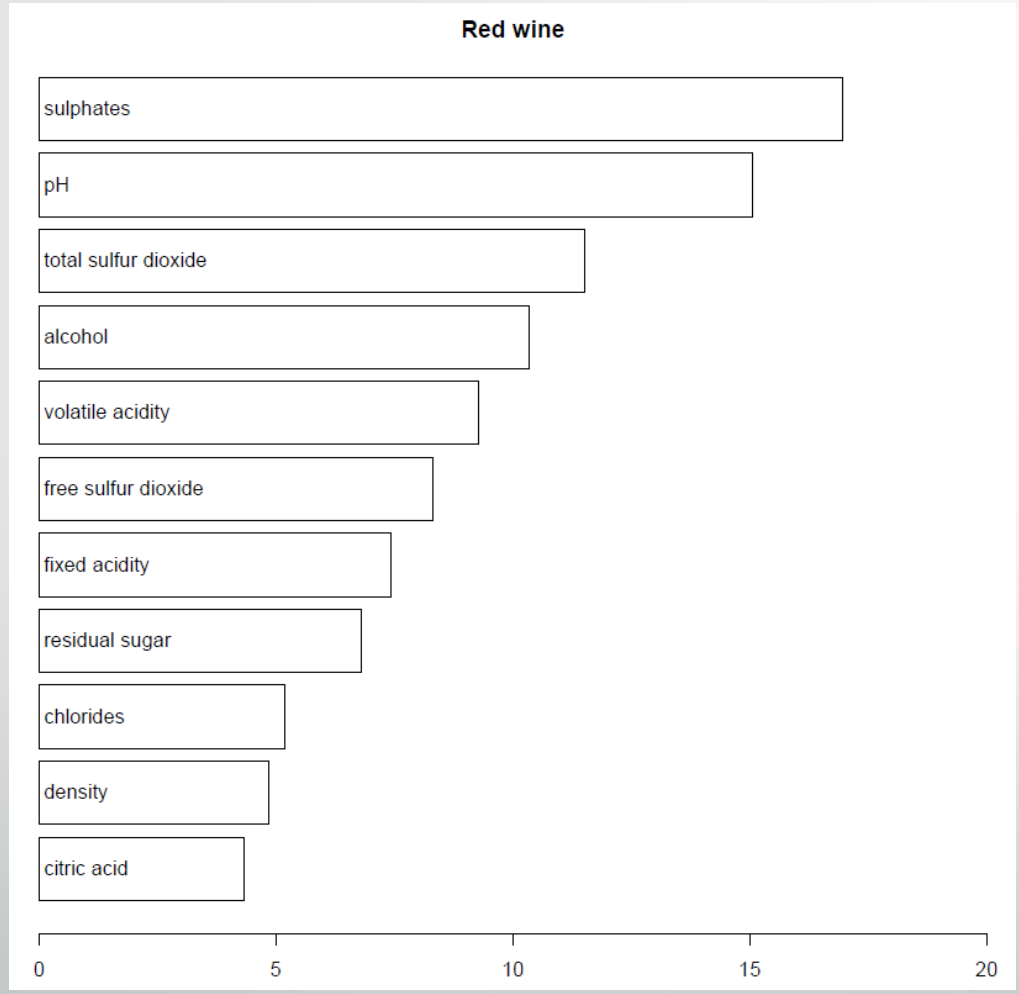


Part2 - Evaluation

- Why accuracy only is about 53%-54%?
  - Quality distribution (K =5) (Training Size: 1088) (Test Size: 271)



Part3 - Conclusions



Inputs importance

Figure references from a relevant same data paper

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.

Part3 - Conclusions

- Algorithm Analysis
  - Data Sampling
    - Training data distribution is too cohesive to be generalized
    - Data set is too small
  - Feature Engineering
    - 11 inputs features are not relatively independence
    - A complex combination of too many features causes overfitting
  - Overfitting
    - Neural Network is not a simple model
    - We cannot observe the learning process within too many inputs, and the outputs may be difficult to explain