Homework 1

The code is submitted at github

<https://github.com/mas-dse/w9yan/blob/master/DSE220/homeworks/homework_2/generative.ipynb>

Answers:

1. Remaining number of rows is 154
2. Removed feature is ‘Ash’ due to >50% missing values

Two features have missing value: ‘Magnesium’, ‘Flavanoids’, after filling

Std(Magnesium) = 14.440377368166187

Std(Flavanoids) = 0.8735732194355235

1. Removed 5 outliers due to sample value for ‘Alcohol’ is 4 standard deviation away from its mean. And from the scatter plot, those 5 outliers have Alcohol < -5 while majority are around 10~15. Intuitively negative value of Alcohol should be mistake.
2. select criterion =  {'gini', 'entropy'}, different leaf sizes were tried as well

Validation accuracy for criterion:gini min\_samples\_leaf:2 - 0.9231

Validation accuracy for criterion:gini min\_samples\_leaf:5 - 0.9231

Validation accuracy for criterion:gini min\_samples\_leaf:10 - 0.8974

Validation accuracy for criterion:gini min\_samples\_leaf:20 - 0.8974

Validation accuracy for criterion:entropy min\_samples\_leaf:2 - 0.9231

Validation accuracy for criterion:entropy min\_samples\_leaf:5 - 0.9487

Validation accuracy for criterion:entropy min\_samples\_leaf:10 - 0.8974

Validation accuracy for criterion:entropy min\_samples\_leaf:20 - 0.8974

Will choose to use criterion: entropy, leaf\_size: 5

test accuracy is: 0.7949

1. Select min samples split={2,5,10,20}

Validation accuracy for min\_samples\_split:2 - 0.9487

Validation accuracy for min\_samples\_split:5 - 0.9487

Validation accuracy for min\_samples\_split:10 - 0.9231

Validation accuracy for min\_samples\_split:20 - 0.9231

Will choose to use min\_sample\_split: 2

test accuracy is: 0.7949

1. using the first 20, 40, 60, 80 and 100 samples from train data.

Validation accuracy with first 20 samples - 0.4615

Validation accuracy with first 40 samples - 0.8462

Validation accuracy with first 60 samples - 0.8462

Validation accuracy with first 80 samples - 0.8974

Validation accuracy with first 100 samples - 0.9487

1. Euclidean distance and k=3

accuracy with Euclidean distance and k=3 is: 0.8718

1. distance metrics defined by l1, linf, l2. And accuracy with best metric and k=3

Validation accuracy for metric manhattan is 0.9487

Validation accuracy for metric euclidean is 0.9231

Validation accuracy for metric chebyshev is 0.9231

Select best metric: manhattan

Test accuracy with metrics=manhattan and k=3 is 0.9744

1. Select k=1,3,5,7,9, and accuracy on selected k

Validation accuracy for k 1 is 0.9487

Validation accuracy for k 3 is 0.9231

Validation accuracy for k 5 is 0.9487

Validation accuracy for k 7 is 0.9744

Validation accuracy for k 9 is 0.9487

Select best k: 7

Test accuracy with Euclidean distance and k=7 is 0.9231

1. Using the first 20, 40, 60, 80 and 100 samples from train data, Euclidean distance and k=3

Validation accuracy with first 20 samples - 0.4615

Validation accuracy with first 40 samples - 0.8462

Validation accuracy with first 60 samples - 0.8462

Validation accuracy with first 80 samples - 0.8974

Validation accuracy with first 100 samples - 0.9487