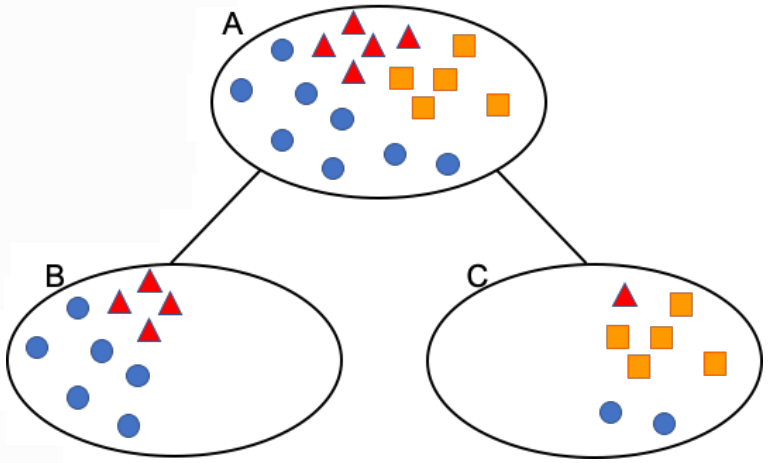
EE2211 Tutorial 9

(Gini impurity, entropy and misclassification rate)

Question 1:

Compute the Gini impurity, entropy, misclassification rate for nodes A, B and C, as well as the overall metrics (Gini impurity, entropy misclassification error) at depth 1 of the decision tree shown below.



RYB total
A 5 5 8 18
GINI A = 1- (5/18)^2 - (5/18)^2 - (8/18)^2 = 0.64814814814

Entropy A = -(5/18)log2(5/18) -(5/18)log2(5/18) -(8/18) log2(8/18) = 1.54663161539

MissClass A = 1 - 8/18 = 0.5555555556

B 4 0 6 10
GINI B = 1- (4/10)^2 - (0/10)^2 - (6/10)^2 = 0.48

Entropy B = -(4/10)log2(4/10) -(0/10)log2(0/10) -(6/10) log2(6/10) = 0.97095059445

MissClass B = 1 - 6/10 = 0.4

C 1 5 2 8
GINI B = 1- (1/8)^2 - (5/8)^2 - (2/8)^2 = 0.53125

Entropy C = -(1/8)log2(1/8) -(5/8)log2(5/8) -(2/8)log2 (2/8) = 1.2987949407

MissClass C = 1 - 5/8 = 0.375

Depth 1
GINI overall = 0.48 * 10/18 + 0.53125 * 8/18 = 0.5027777778

Entropy overall =0.97095059445 * 10/18 + 1.2987949407* 8/18 =1.116659193

miss class overal = 0.4 * 10/18 + 0.375* 8/18 = 0.3888888889

(MSE of regression trees)

Question 2:

Calculate the overall MSE for the following data at depth 1 of a regression tree assuming a decision threshold is taken at $x = 5.0$. How does it compare with the MSE at the root?

$\{x, y\}$: {1, 2}, {0.8, 3}, {2, 2.5}, {2.5, 1}, {3, 2.3}, {4, 2.8}, {4.2, 1.5}, {6, 2.6}, {6.3, 3.5}, {7, 4}, {8, 3.5}, {8.2, 5}, {9, 4.5}

(Regression tree, Python)

Question 3:

Import the California Housing dataset "`from sklearn.datasets import fetch_california_housing`" and "`housing = fetch_california_housing()`". This data set contains 8 features and 1 target variable listed below. Use "`MedInc`" as the input feature and "`MedHouseVal`" as the target output. Fit a regression tree to depth 2 and compare your results with results generated by "`from sklearn.tree import DecisionTreeRegressor`" using the "squared error" criterion.

```
Target: ['MedHouseVal']
Features:['MedInc', 'HouseAge', 'AveRooms', 'AveBedrms', 'Population', 'AveOccup', 'Latitude', 'Longitude']
```

(Classification tree, Python)

Question 4:

Get the data set "`from sklearn.datasets import load_iris`". Perform the following tasks.

(a) Split the database into two sets: 80% of samples for training, and 20% of samples for testing using `random_state=0`

(b) Train a decision tree classifier (i.e., "`tree.DecisionTreeClassifier`" from sklearn) using the training set with a maximum depth of 4 based on the "entropy" criterion.

(c) Compute the training and test accuracies. You can use `accuracy_score` from `sklearn.metrics` for accuracy computation

(d) Plot the tree using "`tree.plot_tree`".