Data Mining, Project 2 – Classification

劉俊毅, N16064103

---

1. Introduction of dataset

1.1 Description

The problem I set for this classification project is *"Whether the scooter drivers of NCKU will run through yellow light or not",* which sounds like a purely fun problem. However, I spent some time thinking of features and trying to generate reasonable values. The following are descriptions of the 7 features and the label of this problem.

Current speed: The speed of the scooter in km per hour when the rider sees the yellow light.

Distance from intersection: The distance of the scooter from the intersection (meter) when the rider sees the yellow light.

Years of riding: How many years the rider has been ridden scooter.

Where from (_TW): Home city of the student - "South", "Middle", "North" and "East" area of Taiwan.

Gender: The physical gender of the rider.

Having backseat passenger: True if the rider is taking a passenger on the backseat, false if not.

Wearing headphone: True if the rider is wearing a headphone, false if not.

Label: 1 = the rider passes the intersection after seeing yellow light

The size of the dataset is not large, only 100 rows. And for the convenience of data generation in Excel with random function, I use specific ranges and intervals for some continuous type of features, e.g. speed and distance … the all the data with features and labels is in "*YellowLight.csv*".

1.2 The rules of label generation

There are number of rules I used for the "absolutely right" property of the labels.

- When the speed is large enough and the distance is short enough, the rider would definitely pass the intersection.
- If the students are from north Taiwan and the riding experience are less than 3 yrs, they would break and stop at the intersection.

- If the students are from south or east Taiwan and their riding experience are more than 3 yrs, they would pass the intersection.
- If the riders are wearing headphones and the distance are short enough, they would pass through.
- If the rider is female and she is taking other in the backseat, she would stop at the intersection.

For the rest of the data, I personality decided its label mostly based on the rules used above. See table1 for the demonstration of some rows of the dataset. See also fig. 1 for the count and pie chart of the two labels.

Table1 A demo for dataset with a few rows

Current speed	Distance from intersection	Years of riding	From (_ TW)	Gender	Have backseat passenger	Wearing headphone	Label
60	25	0	South	F	FALSE	FALSE	1
50	15	0	Middle	M	FALSE	TRUE	1
65	25	2	North	M	FALSE	FALSE	0
45	15	1	Middle	M	TRUE	FALSE	1

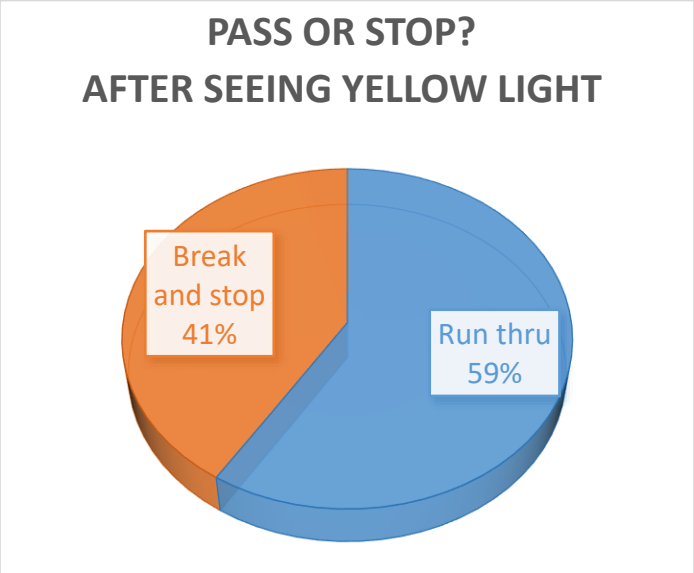

Fig.1 The pie chart of the total count of labels

2. Classification model

- **Decision tree**

See fig. 2 for the brief steps of how the decision tree works. Also, two of the measures of node impurity are used in the classifier in my code, which are GINI index and Entropy. Their computational formulas are shown in eq.1 and 2.

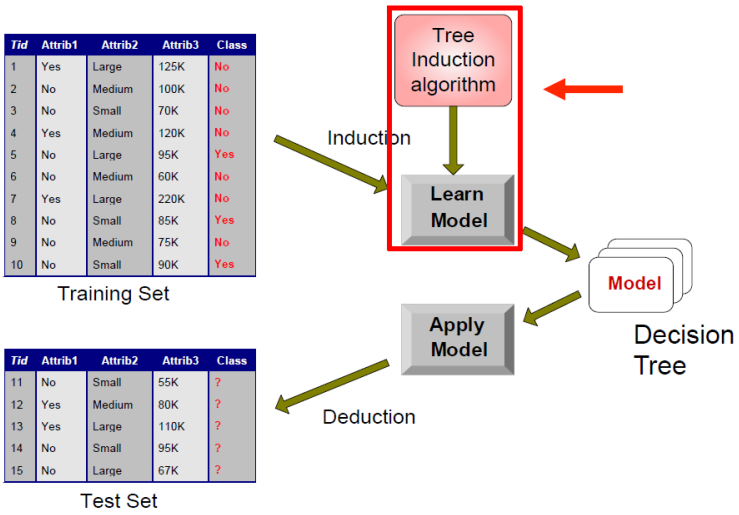

Fig.2 Brief step of decision tree

$$GINI(t) = 1 - \sum [p(j|t)]^2 \qquad \text{Eq. (1)}$$

$$Entropy(t) = -\sum_j p(j|t) \log p(j|t) \qquad \text{Eq. (2)}$$

3. Results and Discussion

3.1 Decision tree

Here, one of the parameters of DecisionTreeClassifier() from sklearn.tree called max_depth is set to be 3 for the clearer evaluation of results, the different value set would be talked about subsequently. See fig. 3 for the plot of decision tree of GINI index. The correspondence table 2 of X to features are listed below.

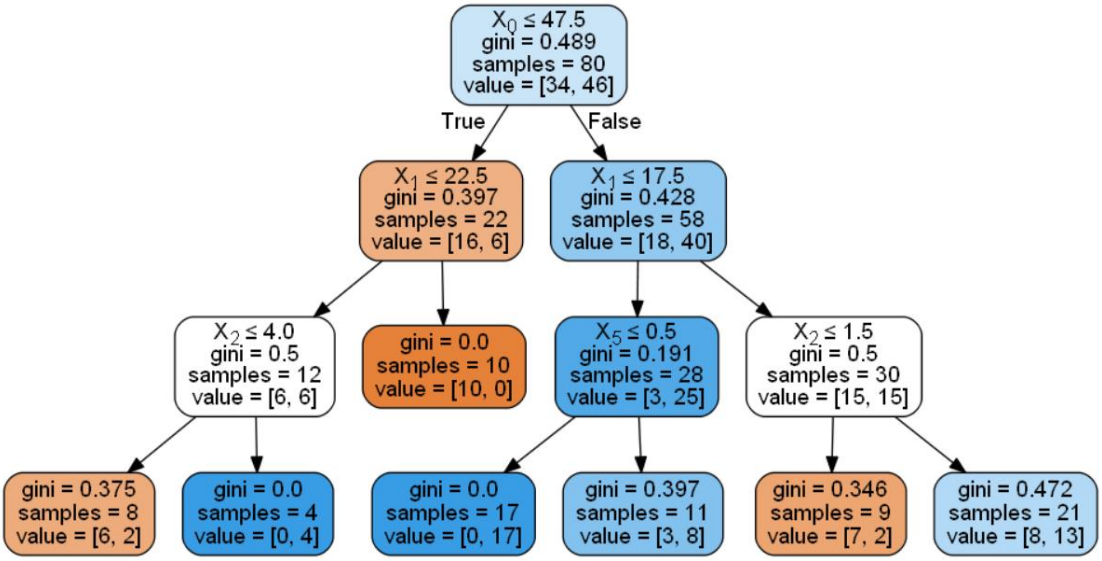

Fig.3 Plot of decision tree using GINI index

Table 2 Correspondence of X in decision tree

Xn	Features
X0	Current speed
X1	Distance from intersection
X2	Years of riding
X3	From which area of TW
X4	Gender
X5	Have backseat passenger
X6	Wearing headphone

For this maximum depth setting (three) of classifier, the decision tree using GINI index and Entropy are basically the same except the value of GINI and entropy calculated, so I would not show the tree plot of Entropy.

The 100-data of the entire dataset is divided into training set (80) and test set (20), the classification model of decision tree is then used to predict the label of test set. In the code, the accuracy score can be calculated to check the performance of the classifier. See table 3 for the accuracy rate of the prediction with the real test set. Like the plot of decision tree, the accuracy scores of GINI index and Entropy are also the same.

Table 3 Accuracy of prediction and real test set using GINI index and Entropy

	GINI index	Entropy
Accuracy score	0.65	0.65

3.2 Comparison with real right data

From the decision tree above, the rules in the decision tree can be compared to the "absolutely right" rules I applied to decide the labels with values of features.

First, the classifier separated the data using the feature of speed, and that's exactly one of my first using feature, which is the fast enough speed along with short enough distance. For the next layer, the distance came into use, which is the X1. At the nodes on the third layer, years of riding and whether having backseat passenger are used. And the feature - years of riding is used in my second and third separation, with the feature of area. Before my subjective judge of label of the rest data, the backseat passenger feature was used with gender – if the rider is female and has a backseat passenger, she would stop before the intersection.

From the decision tree plotted above, however, the area where riders from, their genders and whether they are wearing headphones are **not used**. And these were used in my decision of labels of data.

3.3 Parameters of classifier

- Same dataset split every time

  If the parameter – random_state=… is not used in the split of training and test set from the dataset, the results of data split will be different. So, I use this parameter to make sure the split of dataset is the same every time. (e.g. same test set for every execution of code)

- Depth of decision tree

  It can be understood that if the depth of the decision tree is enough, which means more features may be used to build classification model. So, I tried change the maximum depth from 3 to 5. See fig. 4 and 5 for decision tree using GINI index and Entropy. Obviously, they look more different now. More features were used than that of shallower tree. Also, which can be estimated, the accuracy of the prediction from the model score higher for both GINI and Entropy, see table 4. In this case, the model using Entropy is a better classifier.

Table 4 Accuracy of prediction and real test set using GINI index and Entropy

Deeper tree	GINI index	Entropy
Accuracy score	0.7	0.8

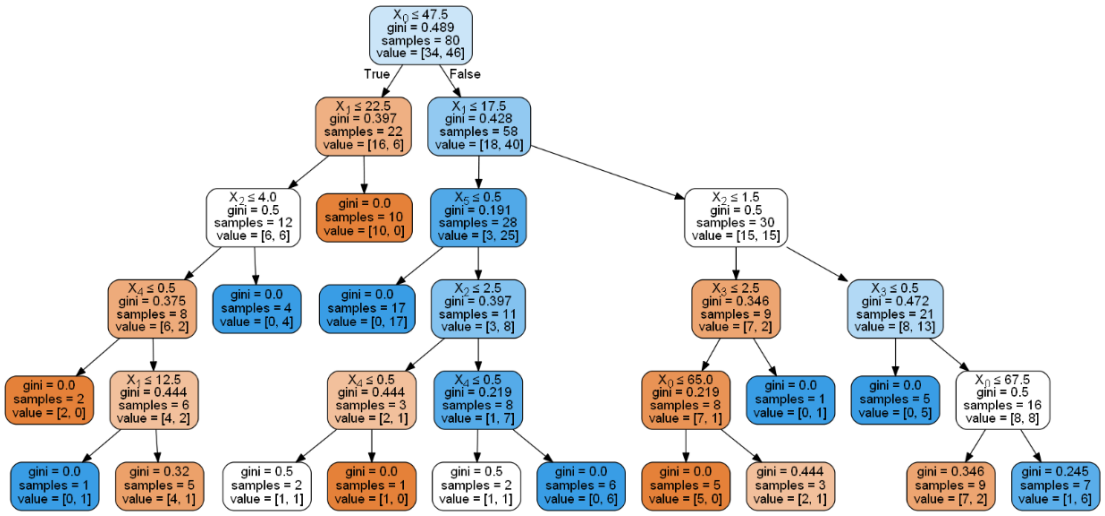

Fig. 4 Plot of decision tree using GINI index (deeper layer)

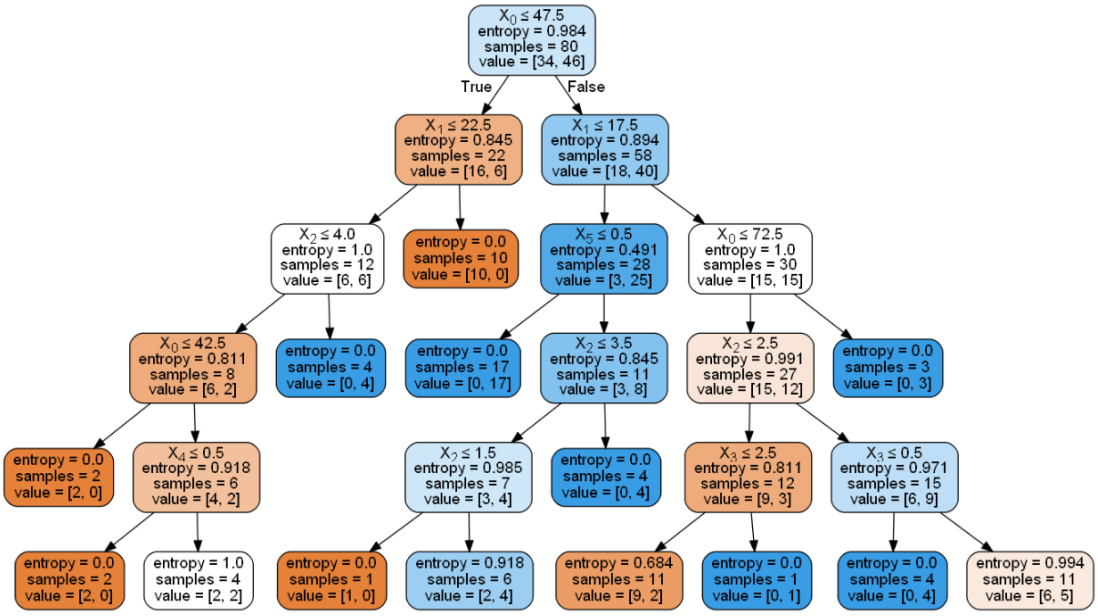

Fig. 4 Plot of decision tree using Entropy (deeper layer)

4. Conclusion

The **size** of the dataset I use may be a bit too small, if the number of data is more than say 1k, perhaps the split of the training and test set will not affect the results of classification even randomly split.

I think my **"absolutely right" rules** are not used so adequately. In excel, I didn't separate the data into two groups in every condition (node), instead, labels of some data is set if they satisfy the condition, and the rest are going into next judging condition. Maybe the way which classifying data into two groups is the better application of "absolutely right" rule when deciding the labels of dataset.

When the depth of the decision tree is **deeper** enough that all or most

of the labels are used in the tree, the accuracy is higher than when some features are not used in the classification model.