Intro. Artificial Intelligent
International University – VNU HCM
Dr. Ly Tu Nga

Course: IT097IU
Date: 03 April 2019
Time: 4 hours

Laboratory Session 3

Decision Tree

1. Data mining

- Data mining refers to extracting knowledge from a large amount of data, in the other way we can say data mining is the process to discover various types of pattern that are inherited in the data and which are accurate, new and useful
- Data mining sometimes called knowledge discovery from data (KDD)

Figure 1 Data mining process

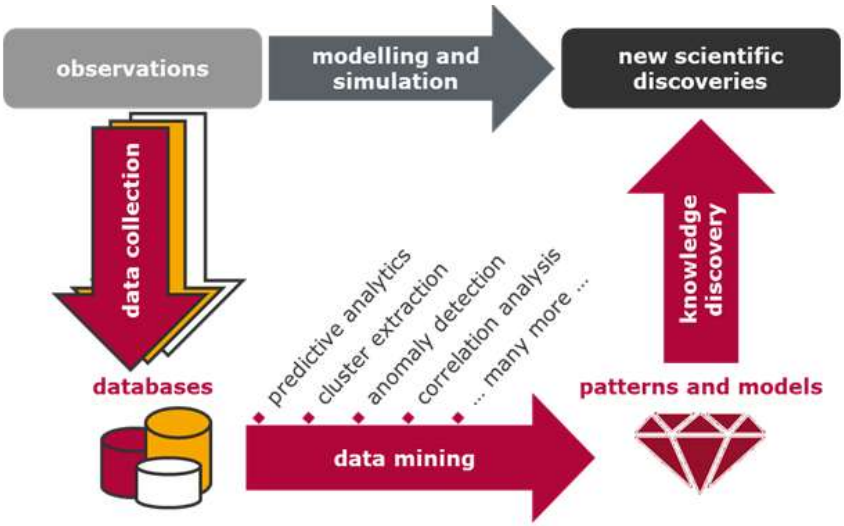
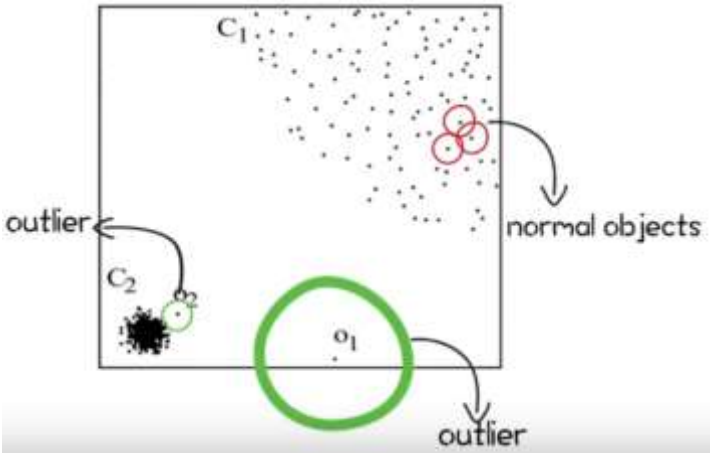
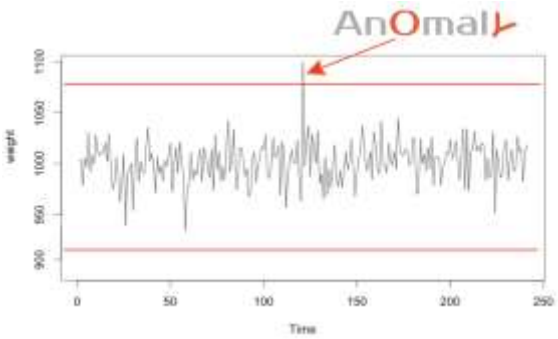


Figure 2

Anomaly and Outlier

- Anomaly detection (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset.
- Most common causes of outliers on a data set:
  - Data entry errors (human errors)
  - Measurement errors (instrument errors)
  - Experimental errors (data extraction or experiment planning/executing errors)
  - Intentional (dummy outliers made to test detection methods)
  - Data processing errors (data manipulation or data set unintended mutations)
  - Sampling errors (extracting or mixing data from wrong or various sources)
  - Natural (not an error, novelties in data)





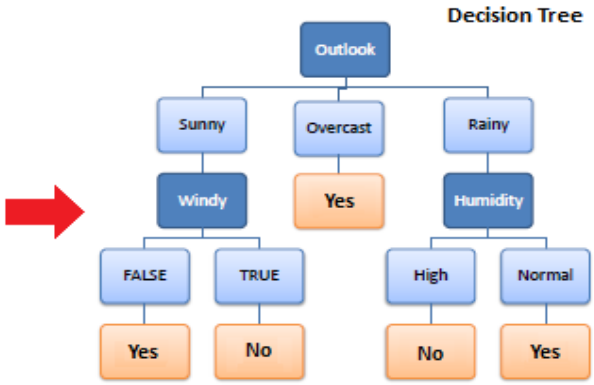
2. Decision Tree [https://scikit-learn.org/stable/modules/tree.html]

Algorithms of DT is https://scikit-learn.org/stable/modules/tree.html#tree-algorithms

- Decision Trees were invented more than 70 years ago, and nowadays they are among the most powerful Machine Learning tools. Main advantage of DTs is that they are a "white-box" method. It means that we can easily explain their decisions, in contrast to the Neural Networks whose complexity is usually too high.

- Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

Predictors				Target
Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



- Splitting criteria

Regression: residual sum of squares

$RSS = \sum_{left} (y_i - y_L^*)^2 + \sum_{right} (y_i - y_R^*)^2$

where y_L^* = mean y-value for left node

y_R^* = mean y-value for right node

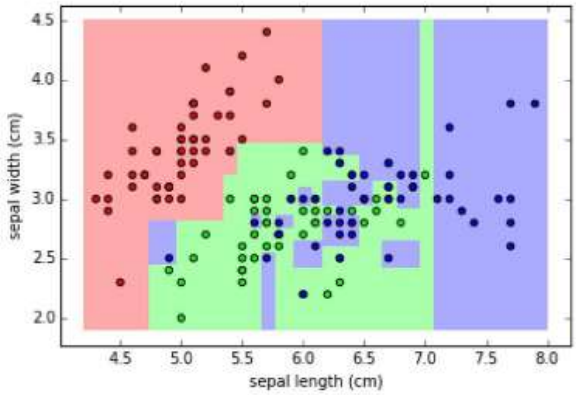
Classification: Gini criterion

$Gini = N_L \sum_{k=1,\ldots,K} p_{kL} (1- p_{kL}) + N_R \sum_{k=1,\ldots,K} p_{kR} (1- p_{kR})$

where p_{kL} = proportion of class k in left node

p_{kR} = proportion of class k in right node

If we keep splitting we will be reducing RSS

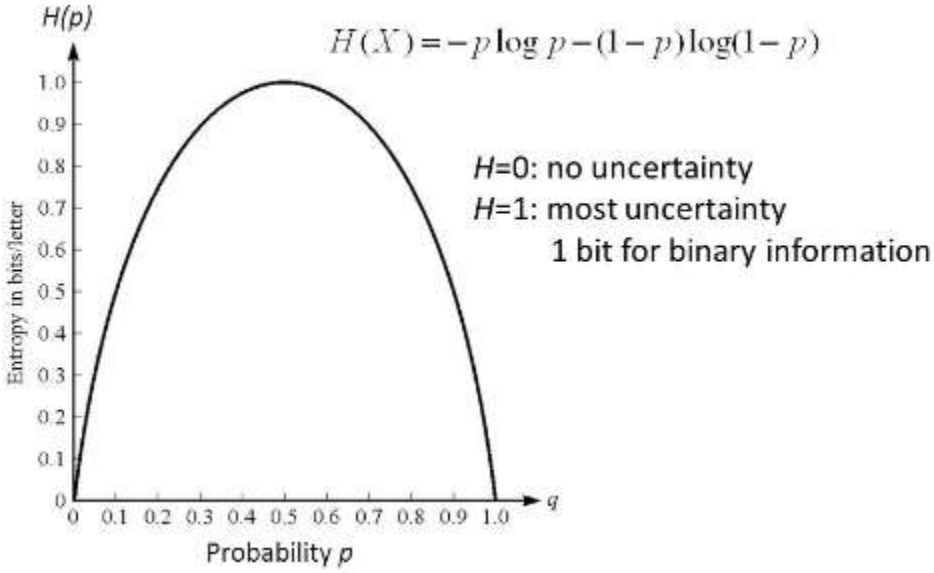


Entropy

- Entropy controls how a Decision Tree decides to split the data. It actually effects how a Decision Tree draws its boundaries.
- The Equation of Entropy:

$$Entropy = -\sum p(X)\log p(X)$$

here p(x) is a fraction of examples in a given class



Information Gain

- The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain.

$$Gain(T,X) = Entropy(T) - Entropy(T,X)$$

- The Equation of Information gain:

Information gain = entropy (parent) – [weightes average] ★ entropy (children)

3. Practice

3.1 Learning without data mining

-read data set “CustomerChurn.csv”

- Learning Process with Encoding for ‘State’, ‘International plan’, ‘voice mail plan’

-Use DataFrame.sample Functions [https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.sample.html] with 2 parameters as frac=TrainDataRate and random_state =1

-Create testing using DataFrame.loc [https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.loc.html] with return a Boolean array where the index values are in values [use: index.isin() https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Index.isin.html]. Therefore: gets the left our portion of the dataset is
Testing=DataFrame.loc[~DataFrame.index.isin(training.index)]

x_train is defined the training dropped TargetAtt as
x_train=training.drop(TargetAtt,1)
y_train is defined the training of TargetAtt as
y_train=training.drop[[TargetAtt]]

x_test is defined the testing dropped TargetAtt as
x_test=testing.drop(TargetAtt,1)
y_train is defined the testing of TargetAtt as
y_test=testing.drop[[TargetAtt]]

+Call SplitDataFrameToTrainAndTest with arguments: DataFrame is data_encoder, TrainDataRate=0.6 and TargetAtt='Churn'
+Print out the values of Train rows and columns; Test rows and columns

+Define the function of Naïve Bayes Learning and Testing as follows

-Print out the value accuracy NBModel with x_train is data_train_encoder, ….

-Define the function of Decision Tree Learning and Testing as

-Print out the value accuracy DT with x_train is data_train_encoder, ….
-Use graphviz to draw the DT

3.2Learning with Data mining
Tutorial ways to detect and remove the outlier is [https://towardsdatascience.com/ways-to-detect-and-remove-the-outliers-404d16608dba]
Finding outliers in dataset [https://medium.com/datadriveninvestor/finding-outliers-in-dataset-using-python-efc3fce6ce32]
Please define the function of DetectOutlierByIQR as follows:

Please define the function of DetectOutlierByLOF and RemoveRowFromDataFrame as follows:

*Learning process with Data Mining

-Create checkOutlierAtt with 3 columns 'Total day minutes', 'Total day calls', 'Total day charge'. And call DetecOutlierByLOF function for this checkOutlierAtt

-Create checkOutlierAtt with 3 columns ‘Total eve minutes’, ‘Total eve calls’, ‘Total eve charge’. And call DetecOutlierByLOF function for this checkOutlierAtt

-Create checkOutlierAtt with 3 columns ‘Total eve minutes’, ‘Total eve calls’, ‘Total eve charge’. And call DetecOutlierByLOF function for this checkOutlierAtt

-Create checkOutlierAtt for ‘Account length’ and call DetectOutlierByIQR function as
-Create checkOutlierAtt for ‘Number email messages’ and call DetectOutlierByIQR function as
- Create checkOutlierAtt for ‘Customer service calls’ and call DetectOutlierByIQR function as

-creat a set and list as
OutlierIndex = set(LOFOutlierIdx01 + LOFOutlierIdx02 + LOFOutlierIdx03 + LOFOutlierIdx04 + IQROutlierIdx01 + IQROutlierIdx02 + IQROutlierIdx03)
OutlierIndex = list(OutlierIndex)
+print out the length of OutlierIndex and string of it.

*create a data_encoder_mining as

- Split Train and Test and check shape

-Retest all traditional classification approaches

And DT as

-Use graphviz to draw the DT