MODULE 4

Prof. Mohammed Tanzeem Agra

MODULE 4 – CONTENTS

1. DECISION TREE
2. REGRESSION
3. ANN
4. CLUSTER ANALYSIS
5. ASSOCIATION RULE

DECISION TREE

- Concept of DT,
- Importance,
- construct,
- algorithm

DECISION TREES

- Decision Tree are a simple way to guide one's path to a decision. The decision may be a simple binary one or it may be a complex multi-valued decision.
- Machine learning can be trained to learn from the past data points and extract some knowledge or rules from it.
- Decision trees use machine learning concepts to abstract knowledge from the data.
- The more data available for training the decision tree, the more accurate its knowledge extraction will be and it will make more accurate decision.
- The more variable the tree can choose from, the greater is the accuracy of the decision tree.

DECISION TREE

- A good DT should also be frugal so that it take the least number of questions and least amount of effort to get the right decision .

DAYS	OUTLOOK	TEMP	HUMIDITY	WINDY	PLAY
1	SUNNY	HOT	HIGH	FALSE	NO
2	SUNNY	HOT	HIGH	TRUE	NO
3	OVERCAST	HOT	HIGH	FALSE	YES
4	RAINY	MILD	HIGH	FALSE	YES
5	RAINY	COOL	NORMAL	FALSE	YES
6	RAINY	COOL	NORMAL	TRUE	NO
7	OVERCAST	COOL	NORMAL	TRUE	YES
8	SUNNY	MILD	HIGH	FALSE	NO
9	SUNNY	COOL	NORMAL	FALSE	YES
10	RAINY	MILD	NORMAL	FALSE	YES
11	SUNNY	MILD	NORMAL	TRUE	YES
12	OVERCAST	MILD	HIGH	TRUE	YES
13	OVERCAST	HOT	NORMAL	FALSE	YES
14	RAINY	MILD	HIGH	TRUE	NO

DECISION TREE CONSTRUCTION

- Determining the Root Node of the Tree
- Splitting the tree

DECISION TREE ALGORITHM

- DECISION TREE APPLY DIVIDE AND CONQUER METHOD.
- PSEUDO CODE FOR MAKING DECISION TREES IS AS FOLLOWS

✓ **CREATE THE ROOT NODE AND ASSIGN ALL THE TRAINING DATA TO IT.**

✓**SELECT THE BEST SPLITTING ATTRIBUTE ACCORDING TO CRITERIA.**

✓**ADD A BRANCH TO THE ROOT NODE FOR EACH VALUE OF THE SPLIT.**

✓**SPLIT THE DATA INTO SUBSET.**

✓**REPEAT STEP 2 AND STEP 3 FOR EACH AND EVERY LEAF NODE .**

DT ALGORITHM DIFFERON TREE KEY ELEMENTS.

- Splitting Criteria
- **Which variable is to use for first split.**
- **Algorithm uses different measures like least errors, gain information, Gini's coefficient.**
- **What values to use for the split?**
- **How many branches should be allowed for each node.**

- Stopping Criteria
- **When to stop building the tree, there are two ways**
- **The tree building can be stopped when a certain dept of branches has been reached and the tree become unrechable.**
- **When the error is in predefined levels.**

DT ALGORITHM DIFFERON TREE KEY ELEMENTS.

- Prunning
- **It is a act of reducing the size of the decision tree by removing section of the tree which gives less information.**
- **When tree contains large amount of data, tree suffer from over- fitting problem.**
- **Preprunning : halt the tree construction early, when certain criteria are met.**
- **Post prunning : removing sub branches or sub tree from fully grown tree.**
- The most popular decision tree algorithm are c5, CART and CHAID

4.2 : REGRESSION

- Regression is a well-known statistical technique to model the predictive relationship between several independent variable(DV) and on dependent variable.
- The objective is to find best fitting – curve for the dependent variable.
- The curve could be a straight line or could be non-linear curve.
- The quality of the data can be measured by a coefficient of a correlation.

The key steps for regression are

1) List all the varaible available for making the models.
2) Establish a Dependent Variable (DV) of interest.
3) Examine Visual relationship between variable of interest.
4) Find a way to predict DV using other variables.

5) There are two types of regression
6) Linear Regression and Non Linear Regression

Advantages and Disadvatages

- Easy to understand as they are built upon statistical principles such as correlation and least square error.
- Provide simple algebraic equations that are easy to understand.
- Regression models can match and beat the predictive power of other modeling tech.
- RM tools are simple. They are found in statistical package as well as data mining package.
- MS-Excel sheet can also be used.

- We cannot cover for data quality issue.
- Regression model suffer from collinearity problem.
- Regression model do not automatically take care of nonlinearity .The user needs to imagine the kind of additional terms that might be needed to improve regression model.
- It work only with numerical data not with categorical data

ASSOCIATION RULE MINING (ARM)

- ARM is a popular, unsupervised learning tech, used in businesses to identify shopping patterns.
- It is also known as "Market Basket Analysis"
- All data used in this tech is of categorical type.
- This tech accepts the raw point-of-sale transaction data as input.The output produced is the description of the most frequent affinities among the items.
- Example : A customer who bought flight tickets and hotel reservation also bought a rental car 60 percent of the time.

REPRESENTING ASSOCIATION RULES

- Can represent set X and Y:
- **X => Y [S%, C%]**
- Where **X,Y** : Product or Services
- **X** : Left Hand Side
- **Y** : Right Hand Side
- **S** (**Support**) : X and Y go together in the dataset i.e **P(XUY)**
- **C** (Confidence) : **Y** is found, given **X**, i.e **P(Y|X)**

REPRESENTING ASSOCIATION RULES (EXAMPLE)

- Suppose there are 1000 transaction in a dataset.
- There are 300 occurrences of X
- There are 150 occurrences of (X,Y)
- Support S for X=>Y will be **P(XUY) = 150/1000 = 15%**
- Confidence for X =>Y will be P(Y|X) or **P(XUY)/P(X) = 150/300 = 50%**

ASSOCIATION RULES EXERCISE (APRIORI ALGORITHM)

TRANSACTION LIST				
1	MILK	EGG	BREAD	BUTTER
2	MILK	BUTTER	EGG	KETCHUP
3	BREAD	BUTTER	KETCHUP	
4	MILK	BREAD	BUTTER	
5	BREAD	BUTTER	COOKIES	
6	MILK	BREAD	BUTTER	COOKIES
7	MILK	COOKIES		
8	MILK	BREAD	BUTTER	
9	BREAD	BUTTER	EGG	COOKIES
10	MILK	BUTTER	BREAD	
11	MILK	BREAD	BUTTER	
12	MILK	BREAD	COOKIES	KETCHUP