MODULE 06

DATA SCIENCE BOOTCAMP

Supervised Learning - Classification Models

Vertical Institute

Reminder

1. Please turn on Zoom camera the whole duration of classes.

2. At the start of all classes, please rename yourselves to: Name + Last 3 digits and letter of your NRIC. Example: John Tan (123A)

Agenda

- K-Nearest Neighbors
- Logistic Regression
- Decision Tree Model
- Model Evaluation

Attendance Photo Taking

MODULE 6: SUPERVISED LEARNING - CLASSIFICATION MODELS

K-Nearest Neighbors

What is Classification Model?

Classification is used when the target outputs are discrete.

Classification model requires:

- Training, validation and testing data
- Input features (X)
- Target features / output (Y)
- Measures of Improvement

What Kind of Questions It Answers?

Is that bank client going to default on the loan?

Will the user who clicked on the ad buy the product?

Who is the person in the Facebook photo?

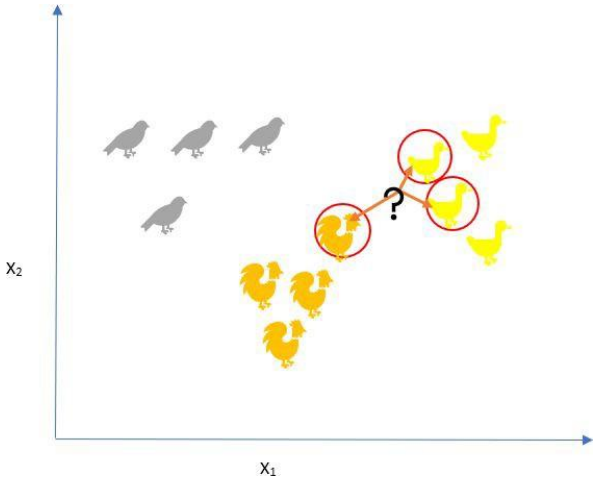
Binary classification problem - if there are only 2 labels, “default / no default”

Multi-class classification problem — If there are more labels, eg moody’s ratings “AAA”/”BBB”/”CCC”/

K-Nearest Neighbours (KNN) Model

What is K-Nearest Neighbour (KNN) Model?

K-Nearest Neighbors (KNN) is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points.

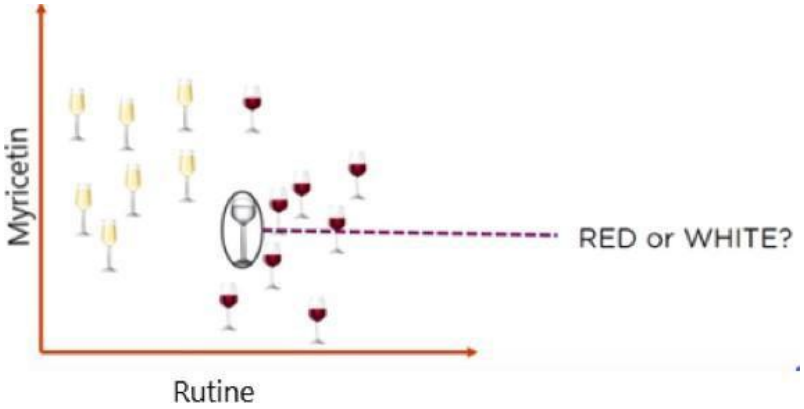


"Birds of a feather flock together"

How It Works?

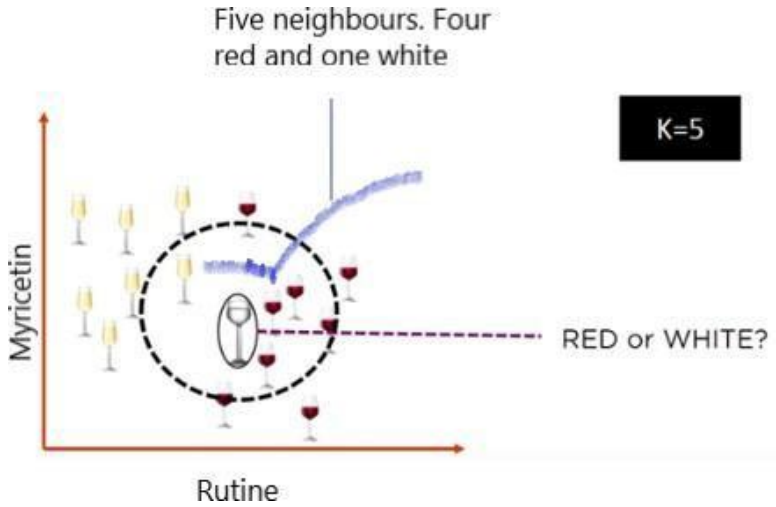
Suppose, if we add a new glass of wine in the dataset, we would like to know whether the new wine is red or white.

So, we need to find out what the neighbours are in this case.



How It Works?

Let’s say k = 5 and the new data point is classified by the majority of votes from its five neighbours and the new point would be classified as red since four out of five neighbours are red.



How To Choose The Value For K?

Choosing the right value of k is a process called **parameter tuning** and is important for better accuracy.
However, finding the value of k is not easy.

Few ideas on picking a value for k:

1. There is no structured method to find the best value for “K”. We need to find out with various values by trial and error and assuming that training data is unknown.
2. Choosing smaller values for K can be noisy and will have a higher influence on the result.
3. Larger values of K will have smoother decision boundaries which mean lower variance but increased bias. Also, computationally expensive.

Features of KNN Model

- **Non-parametric:** does not make assumptions about the underlying distribution for our data.
- **Lazy:** training phase is minimal - KNN uses all (or nearly all) of the training data.
- **Based on feature similarity:** how closely out-of-sample features resemble our training set determines how we classify a given data point.

Advantages/Drawbacks

Benefits

- Simple to understand and explain
- Model training is fast
- Can be used for classification and regression

Drawbacks

- Must store all of the training data.
- Prediction phase can be slow when n is large.
- Sensitive to irrelevant features.
- Sensitive to the scale of the data
- Accuracy is (generally not competitive with other supervised learning models.

Train — Test - Split

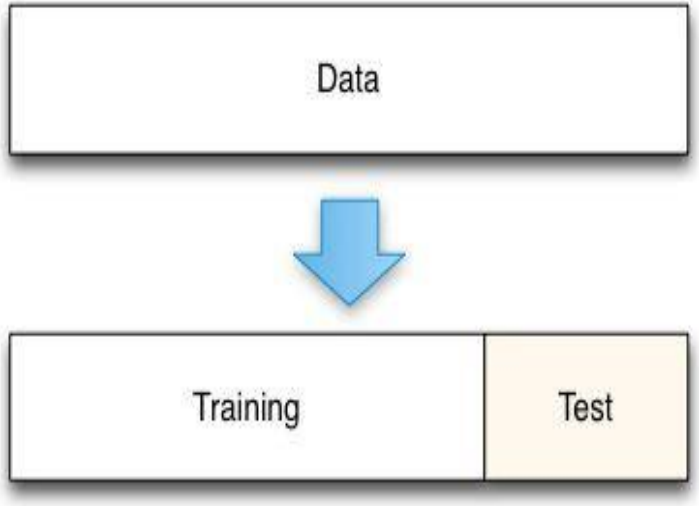
The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model.

The objective is to estimate the performance of the machine learning model on new data: data not used to train the model.

Train — Test - Split

The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the **training dataset**.

The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the **test dataset**.



MODULE 6: SUPERVISED LEARNING - CLASSIFICATION MODELS

Logistic Regression Model

What is Logistic Regression Model?

Considering the scenario:

1. Email is spam or not.
2. Will the customer buy life insurance?
3. Which party a person is going to vote for?

Predicted values is categorical

What is Logistic Regression Model?

Considering an insurance dataset:

You have people of different age

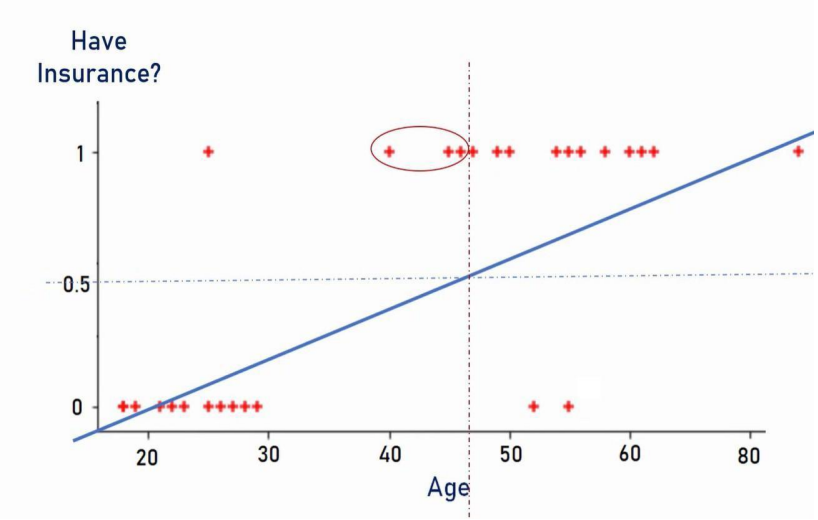
either bought an insurance or not.

age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0

What is Logistic Regression Model?

You will a plot a scatter plot that will look like a regression model.

However, it is not good fit for prediction.



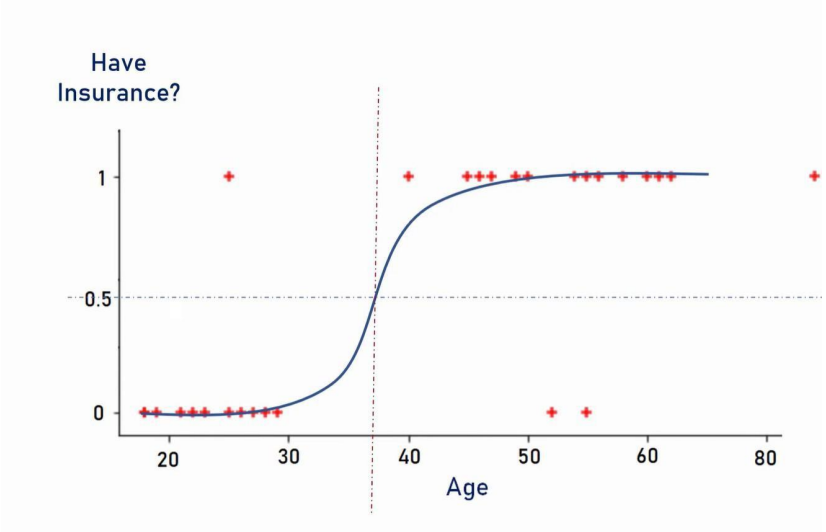
What is Logistic Regression Model?

We use the **activation function** (sigmoid) to convert the outcome into categorical value.

$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

e = Euler's number ~ 2.71828

Sigmoid function converts input into range 0 to 1

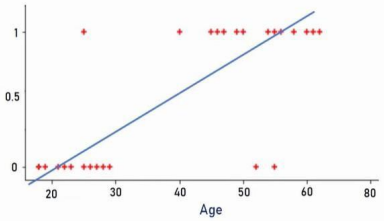


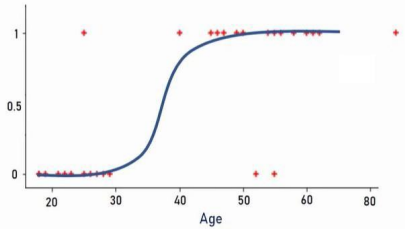
What is Logistic Regression Model?

Essentially, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes.

$$y = m * x + b$$

$$y = \frac{1}{1 + e^{-(m*x+b)}}$$





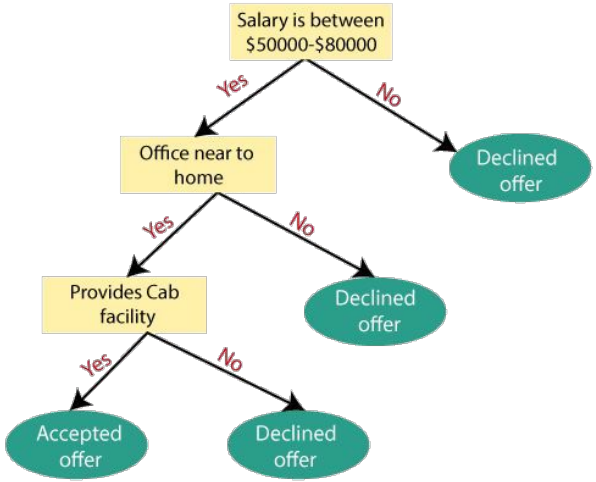
MODULE 6: SUPERVISED LEARNING - CLASSIFICATION MODELS

Decision Tree Model

What is Decision Tree Model?

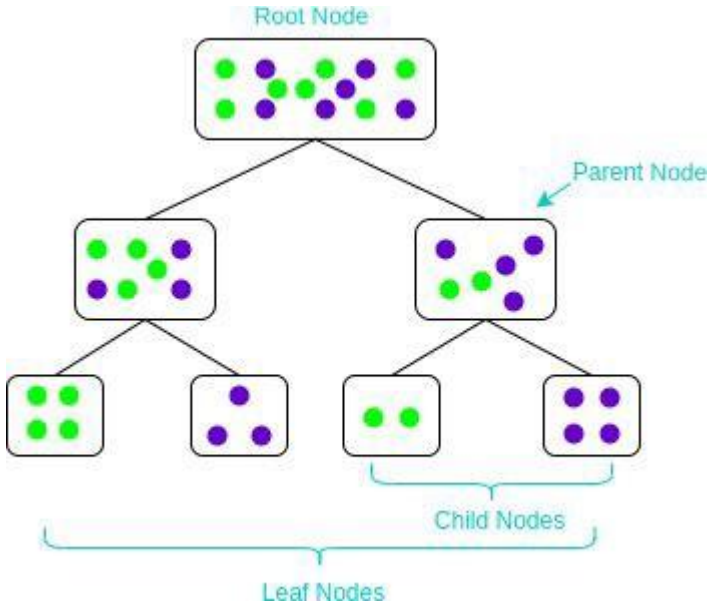
A Decision Tree is a supervised algorithm used in the machine learning. It is using a binary tree graph (each node has two children) to assign for each data sample a target value. Target values are presented in the tree leaves.

Decision Tree learning is a process of finding the optimal rules in each internal tree node according to the selected metrics.



What is Decision Tree Model?

- **Parent and Child Node:** A node that gets divided into sub-nodes is known as Parent Node, and these sub-nodes are known as Child Nodes. Since a node can be divided into multiple sub-nodes, therefore a node can act as a parent node of numerous child nodes.
- **Root Node**: The top-most node of a decision tree. It does not have any parent node. It represents the entire population or sample.
- **Leaf/Terminal Nodes**: Nodes that do not have any child node are known as Terminal/Leaf Nodes.



What is Decision Tree Model?

This process of classification begins with the root node of the decision tree and expands by applying some splitting conditions at each non-leaf node, it divides datasets into a homogeneous subset.

The ‘knowledge’ learned by a decision tree through training is directly formulated into hierarchical structure.

In order to check “the goodness of splitting criterion” or for evaluating how well the splitting is, various splitting indies were proposed. Some of them are **gini index** and **information gain.**

Advantages/Disadvantages

Advantages

- Decisions are easy to understand and interpret.
- Weight and importance of each feature becomes clear.
- Numerical and categorical features can be used naturally.
- Trees are a natural multi-class classifier.

Disadvantages

- Overfit to training data with complex trees.
- Small changes in input data can result in totally different trees.
- Can make mistakes with unbalanced classes.
- Requires large datasets to build robust rule.

Recap time!

What are your favorite takeaways on Classification Models today?

Let's share with each other!

Some things to take note...

Link and resource could be accessed in the Learning Portal.

https://elearn.verticalinstitute.com/users/sign_in

Attendance Photo Taking

MODULE 06

DATA SCIENCE BOOTCAMP

Thank you!

Vertical Institute