LECTURE 8

MACHINE LEARNING

DR. PRAPASSORN TANTIPHANWADI
INDUSTRIAL ENGINEERING, FACULTY OF ENGINEERING AT KHAMPAENGSAEN
DECEMBER 2565

CONTENT

- What is Machine Learning?
- Type of Machine Learning?
- Overfitting and Underfitting
- ML Algorithms
    1) Simple Linear Regression
    2) Multiple Linear Regression
    3) Logistic Regression
    4) Decision Tree
    5) Random Forest
    6) Support Vector Machine (SVM)
    7) Naiive Bayes
    8) K-NN
    9) PCA
    10) K-Mean Clustering

WHAT IS MACHINE LEARNING

Traditional Programming

- Data (Input)
- Program

Output

Machine Learning

- Data (Input)
- Program

Program,
Simulation model

EXAMPLES OF MACHINE LEARNING

1. Image recognition

Image recognition is a well-known and widespread example of machine learning in the real world. It can identify an object as a digital image, based on the intensity of the pixels in black and white images or colour images.

Real-world examples of image recognition:

- Label an x-ray as cancerous or not
- Assign a name to a photographed face (aka "tagging" on social media)
- Recognize handwriting by segmenting a single letter into smaller images

Machine learning is also frequently used for facial recognition within an image. Using a database of people, the system can identify commonalities and match them to faces. This is often used in law enforcement.

EXAMPLES OF MACHINE LEARNING

2. Speech recognition
Machine learning can translate speech into text. Certain software applications can convert live voice and recorded speech into a text file. The speech can be segmented by intensities on time-frequency bands as well.

Real-world examples of speech recognition:

- Voice search
- Voice dialling
- Appliance control

Some of the most common uses of speech recognition software are devices like Google Home or Amazon Alexa.

EXAMPLES OF MACHINE LEARNING

3. Medical diagnosis
Machine learning can help with the diagnosis of diseases. Many physicians use chatbots with speech recognition capabilities to discern patterns in symptoms.

Real-world examples for medical diagnosis:

- Assisting in formulating a diagnosis or recommends a treatment option
- Oncology and pathology use machine learning to recognise cancerous tissue
- Analyse bodily fluids

In the case of rare diseases, the joint use of facial recognition software and machine learning helps scan patient photos and identify phenotypes that correlate with rare genetic diseases.

EXAMPLES OF MACHINE LEARNING

4. Statistical arbitrage
Arbitrage is an automated trading strategy that's used in finance to manage a large volume of securities. The strategy uses a trading algorithm to analyse a set of securities using economic variables and correlations.

Real-world examples of statistical arbitrage:

- Algorithmic trading which analyses a market microstructure
- Analyse large data sets
- Identify real-time arbitrage opportunities

Machine learning optimises the arbitrage strategy to enhance results.

EXAMPLES OF MACHINE LEARNING

5. Predictive analytics
Machine learning can classify available data into groups, which are then defined by rules set by analysts. When the classification is complete, the analysts can calculate the probability of a fault.

Real-world examples of predictive analytics:

- Predicting whether a transaction is fraudulent or legitimate
- Improve prediction systems to calculate the possibility of fault

Predictive analytics is one of the most promising examples of machine learning. It's applicable for everything; from product development to real estate pricing.

EXAMPLES OF MACHINE LEARNING

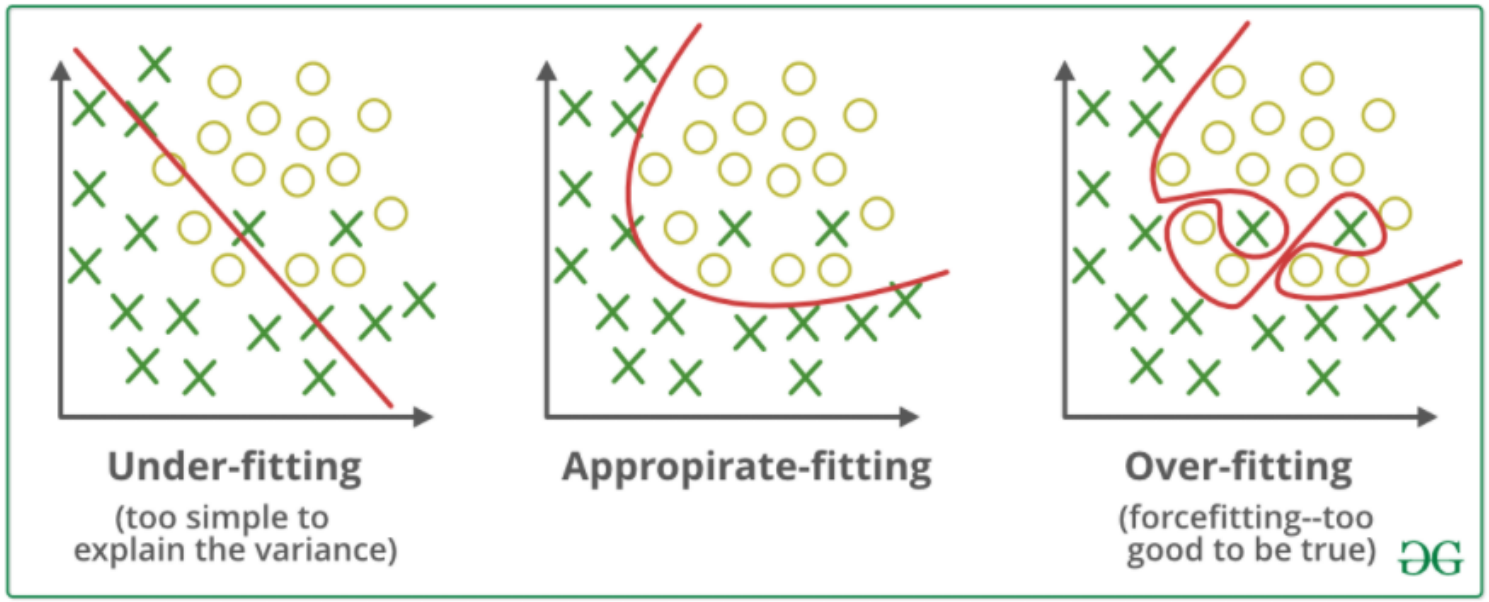
6. Extraction
Machine learning can extract structured information from unstructured data. Organisations amass huge volumes of data from customers. A machine learning algorithm automates the process of annotating datasets for predictive analytics tools.

Real-world examples of extraction:

- Generate a model to predict vocal cord disorders
- Develop methods to prevent, diagnose, and treat the disorders
- Help physicians diagnose and treat problems quickly

Typically, these processes are tedious. But machine learning can track and extract information to obtain billions of data samples.

OVERFITTING AND UNDERFITTING



Techniques to reduce overfitting:

1. Increase training data.
2. Reduce model complexity.
3. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
4. Ridge Regularization and Lasso Regularization
5. Use dropout for neural networks to tackle overfitting.

TYPE OF ML

Supervised learning

Supervised learning, as the name indicates, has the presence of a supervisor as a teacher. Basically supervised learning is when we teach or train the machine using data that is well labeled. Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data(set of training examples) and produces a correct outcome from labeled data.

For instance, suppose you are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this:

- If the shape of the object is rounded and has a depression at the top, is red in color, then it will be labeled as –Apple.
- If the shape of the object is a long curving cylinder having Green-Yellow color, then it will be labeled as –Banana.

TYPE OF ML Supervised learning

- **Supervised learning** is classified into two categories of algorithms:

- **Classification:** A classification problem is when the output variable is a category, such as "Red" or "blue" or "disease" and "no disease".
- **Regression:** A regression problem is when the output variable is a real value, such as "dollars" or "weight". Supervised learning deals with or learns with "labeled" data. This implies that some data is already tagged with the correct answer.

- **Types:-**
  - Regression
  - Logistic Regression
  - Classification
  - Naive Bayes Classifiers
  - K-NN (k nearest neighbors)
  - Decision Trees
  - Support Vector Machine

Advantages:-

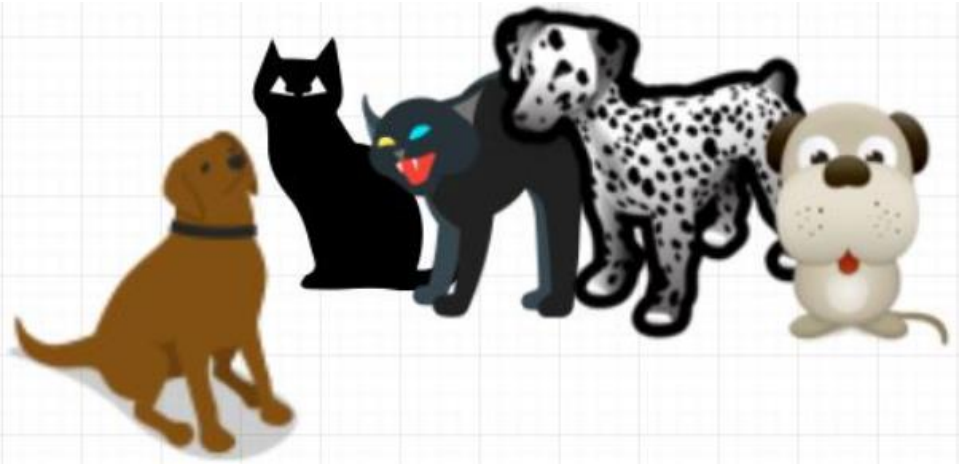
Supervised learning allows collecting data and produces data output from previous experiences. Helps to optimize performance criteria with the help of experience. Supervised machine learning helps to solve various types of real-world computation problems.

Disadvantages:-

Classifying big data can be challenging. Training for supervised learning needs a lot of computation time. So, it requires a lot of time.

TYPE OF ML Unsupervised learning

- **Unsupervised learning** is unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.
- Unlike supervised learning, no teacher is provided that means no training will be given to the machine. Therefore the machine is restricted to find the hidden structure in unlabeled data by itself.
- For instance, suppose it is given an image having both dogs and cats which it has never seen.



- Thus the machine has no idea about the features of dogs and cats so we can't categorize it as 'dogs and cats '. But it can categorize them according to their similarities, patterns, and differences, i.e., we can easily categorize the above picture into two parts. The first may contain all pics having dogs in them and the second part may contain all pics having cats in them. Here you didn't learn anything before, which means no training data or examples.
- It allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with unlabelled data.

TYPE OF ML Unsupervised learning

- **Unsupervised learning** is classified into two categories of algorithms:

- **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

- **Types of Unsupervised Learning:-**

  - **Clustering**
    1) Exclusive (partitioning)
    2) Agglomerative
    3) Overlapping
    4) Probabilistic

  - **Clustering Types:-**
    1) Hierarchical clustering
    2) K-means clustering
    3) Principal Component Analysis
    4) Singular Value Decomposition
    5) Independent Component Analysis

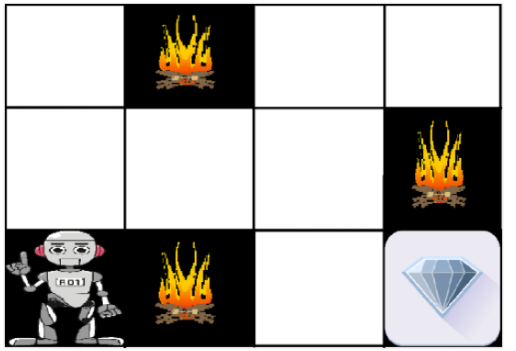
TYPE OF ML

Supervised vs. Unsupervised Machine Learning

Parameters	Supervised machine learning	Unsupervised machine learning
Input Data	Algorithms are trained using labeled data.	Algorithms are used against data that is not labeled
Computational Complexity	Simpler method	Computationally complex
Accuracy	Highly accurate	Less accurate

TYPE OF ML Reinforcement learning

- **Reinforcement learning** is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.



- Example: The problem is as follows: We have an agent and a reward, with many hurdles in between. The agent is supposed to find the best possible path to reach the reward. The following problem explains the problem more easily.

- **Main points in Reinforcement learning –**
  - Input: The input should be an initial state from which the model will start
  - Output: There are many possible outputs as there are a variety of solutions to a particular problem
  - Training: The training is based upon the input, The model will return a state and the user will decide to reward or punish the model based on its output.
  - The model keeps continues to learn.
  - The best solution is decided based on the maximum reward.
  - Difference between Reinforcement learning and Supervised learning:

TYPE OF ML

Difference between Reinforcement learning and Supervised learning:

Reinforcement learning	Supervised learning
Reinforcement learning is all about making decisions sequentially. In simple words, we can say that the output depends on the state of the current input and the next input depends on the output of the previous input	In Supervised learning, the decision is made on the initial input or the input given at the start
In Reinforcement learning decision is dependent, So we give labels to sequences of dependent decisions	In supervised learning the decisions are independent of each other so labels are given to each decision.
Example: Chess game	Example: Object recognition

TYPE OF ML Reinforcement learning

- Types of Reinforcement: There are two types of Reinforcement:

1) **Positive –**
   - Positive Reinforcement is defined as when an event, occurs due to a particular behavior, increases the strength and the frequency of the behavior. In other words, it has a positive effect on behavior.
   - Advantages of reinforcement learning are:
   - Maximizes Performance
   - Sustain Change for a long period of time
   - Too much Reinforcement can lead to an overload of states which can diminish the results

2) ***Negative –***
   - Negative Reinforcement is defined as strengthening of behavior because a negative condition is stopped or avoided.
   - Advantages of reinforcement learning:
   - Increases Behavior
   - Provide defiance to a minimum standard of performance
   - It Only provides enough to meet up the minimum behavior

TYPE OF ML Reinforcement learning

- **Various Practical applications of Reinforcement Learning –**
  - RL can be used in robotics for industrial automation.
  - RL can be used in machine learning and data processing
  - RL can be used to create training systems that provide custom instruction and materials according to the requirement of students.

- **RL can be used in large environments in the following situations:**
  - A model of the environment is known, but an analytic solution is not available;
  - Only a simulation model of the environment is given (the subject of simulation-based optimization)
  - The only way to collect information about the environment is to interact with it.

SCIKIT-LEARN WITH ML

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

https://scikit-learn.org/stable/

SIMPLE LINEAR REGRESSION

$$Y = a+bX$$

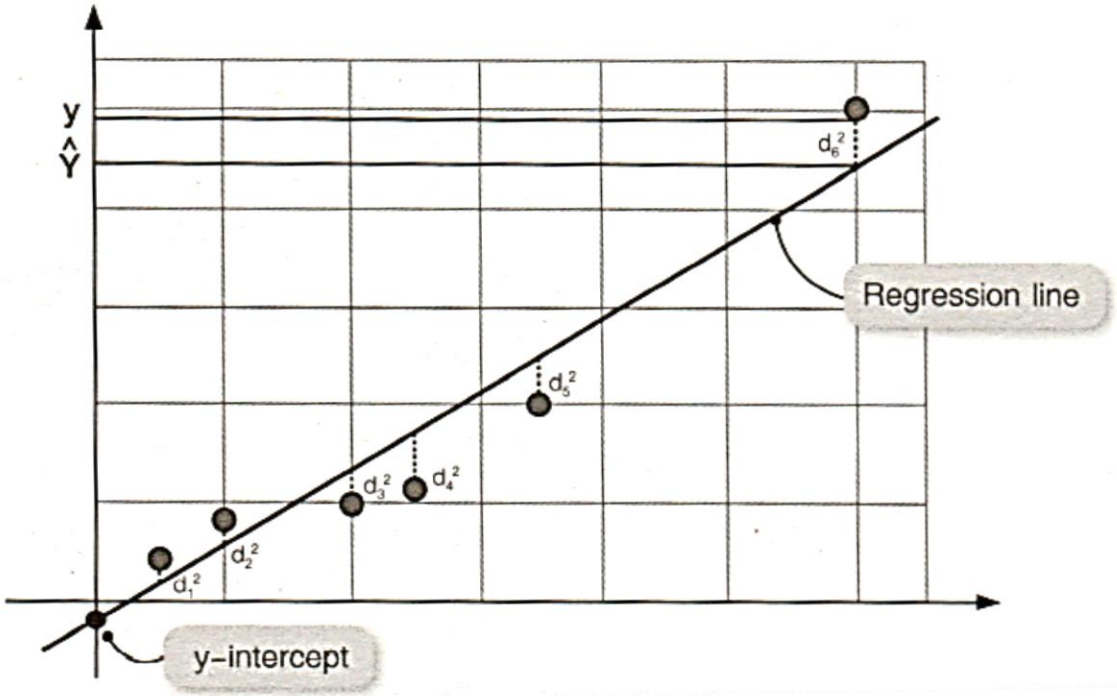
$$a = \bar{Y} - b\bar{X}$$

โดยเส้นประ (แทนด้วยตัวแปร d) คือ ผลต่างระหว่างค่าจริงและค่าทำนาย ซึ่งคำนวณได้จากสูตร

$$d = y - \hat{Y}$$

$$D = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$$

$$b = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum X^2 - \frac{(\sum X)^2}{n}}$$



ค่าโฆษณา (บาท)	ยอดขาย (บาท)
450	1200
600	1500
750	2000
500	1400
650	1550
1000	3500

MULTIPLE LINEAR REGRESSION

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n$$

ประสบการณ์ (ปี)	คะแนน TOEIC	เงินเดือน
1	700	30000
2	350	25000
5	450	38000
3	900	35000
3	500	31000
5	650	42000

HW

LOGISTIC REGRESSION

Logistic Regression คือ การวิเคราะห์การถดถอยโลจิสติก เป็นอัลกอริทึมหนึ่งของ Machine Learning ที่จัดอยู่ในประเภท Supervised Learning

เราสามารถแบ่งการวิเคราะห์ Logistic Regression ออกได้เป็น 2 ประเภท คือ

- **Binary Logistic Regression** คือ การวิเคราะห์โลจิสติกแบบทวิ ซึ่งตัวแปรตามจะมี 2 ค่า คือ เกิดเหตุการณ์ (แทนด้วยค่า y=1) หรือ ไม่เกิดเหตุการณ์ (แทนด้วยค่า y=0)
- **Multinomial Logistic Regression** คือ การวิเคราะห์โลจิสติกแบบพหุกลุ่ม ซึ่งตัวแปรตามจะมีมากกว่า 2 ค่า เช่น แย่ (แทนด้วยค่า y=3) ปานกลาง (แทนด้วยค่า y=2) ดี (แทนด้วยค่า y=1)

การวิเคราะห์โลจิสติกมีเป้าหมาย ก็คือ เพื่อทำนายโอกาสความน่าจะเป็น (Probability) ที่จะเกิดเหตุการณ์ที่สนใจ โดยใช้ตัวแปรอิสระ (ตัวแปร X) 1 ตัวหรือมากกว่า 1 ตัว เพื่อนำมาวิเคราะห์ โดยชนิดข้อมูลของตัวแปรอิสระแบ่งได้เป็น 4 ประเภท คือ

- **Dichotomous** คือ ตัวแปรอิสระที่มีได้ 2 ค่า เช่น เพศ (ชาย-หญิง), ผลการสอบ (ผ่าน-ไม่ผ่าน) เป็นต้น
- **Interval scale** คือ ตัวแปรอิสระที่ข้อมูลมีช่วงวัดที่ห่างเท่ากัน เช่น เกรด (A-B-C-D-F) เป็นต้น
- **Ratio scale** คือ ตัวแปรอิสระที่มีค่าในเชิงตัวเลขที่แท้จริง ไม่ได้สมมติขึ้น เช่น น้ำหนัก ส่วนสูง ระยะทาง ความเร่ง ความเร็ว เป็นต้น
- **Polytomous** คือ ตัวแปรอิสระที่มีค่าแตกต่างกันมากกว่า 2 ค่า เช่น ศาสนา (พุทธ-คริสต์-ฮินดู) เชื้อชาติ (ไทย-จีน-อิสลาม) เป็นต้น

LOGISTIC REGRESSION

ตัวอย่างของการนำอัลกอริทึม Logistic Regression ไปใช้งาน เช่น

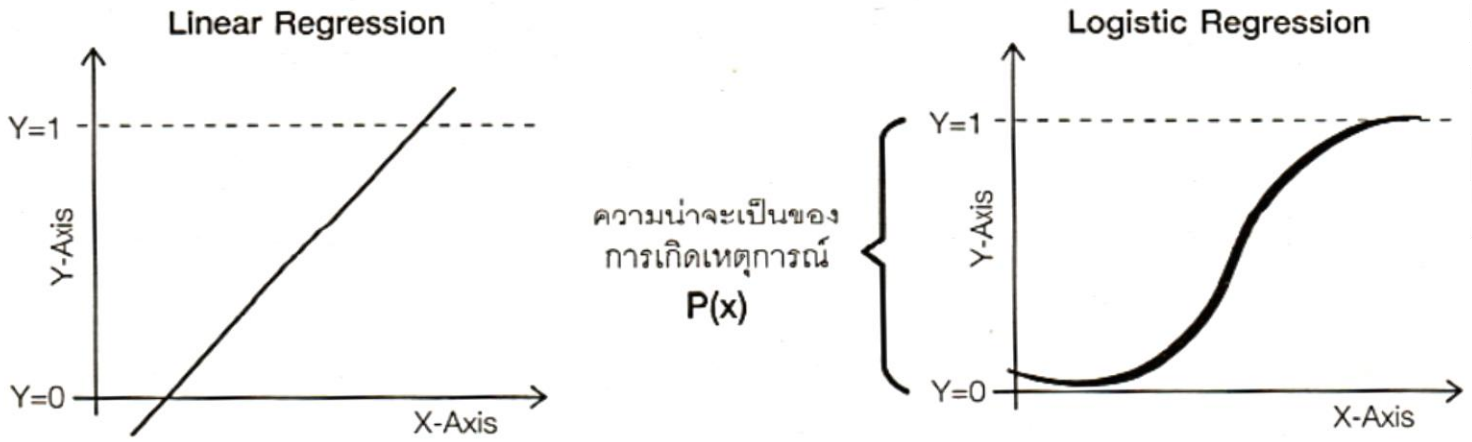
- การทำนายว่าอีเมลมีความน่าจะเป็นที่จะเป็นสแปมหรือไม่ โดยนำ IP Address, หัวข้ออีเมล, ผู้ส่งอีเมลมาเป็นตัวแปรในการวิเคราะห์ คำตอบ คือ ใช่/ไม่ใช่
- การทำนายว่าผู้ป่วยมีความน่าจะเป็นที่จะเป็นโรคเบาหวานหรือไม่ โดยนำระดับน้ำตาลในเลือด, ระดับน้ำตาลสะสมในเลือดย้อนหลัง 2-3 เดือน, ประวัติทางพันธุกรรม, อายุ, เพศ มาเป็นตัวแปรในการวิเคราะห์ คำตอบ คือ เป็น/ไม่เป็น
- การทำนายว่ามีความน่าจะเป็นที่จะอนุมัติเงินกู้ให้กับลูกค้าหรือไม่ โดยนำรายได้ต่อเดือน, สถานะเครดิตบูโร, อายุ, ยอดเงินกู้ มาเป็นตัวแปรในการวิเคราะห์ คำตอบ คือ อนุมัติ/ไม่อนุมัติ

LOGISTIC REGRESSION

เราลองมาดูหลักการของ Binary Logistic Regression กันค่ะ ก่อนอื่นพิจารณาความแตกต่างระหว่าง Linear Regression และ Logistic Regression กันก่อนว่าต่างกันอย่างไร

Linear Regression จะให้ผลลัพธ์ออกมาเป็นค่าที่ต่อเนื่องกัน (continuous) เช่น ราคาบ้าน ราคาหุ้น แต่ Logistic Regression จะให้ผลลัพธ์ออกมาเป็นค่าที่ไม่ต่อเนื่องกัน (discrete) เช่น ผู้ป่วยเป็นโรคหัวใจหรือไม่ ลูกค้าจะกลับมาซื้อสินค้าซ้ำหรือไม่

โดยค่าผลลัพธ์ของ Linear Regression จะเป็นค่าที่ต่ำกว่า 0 หรือมากกว่า 1 ก็ได้ แต่สำหรับ Logistic Regression แล้วตัวแปรตามจะมีได้ 2 ค่า คือ 0 (ไม่เกิดเหตุการณ์) กับ 1 (เกิดเหตุการณ์) เท่านั้น ความสัมพันธ์ของตัวแปรอิสระ x และ ตัวแปรตาม y จึงไม่มีทางที่จะอยู่ในรูปเส้นตรง แต่ความสัมพันธ์ของ x และ y จะเป็นรูปคล้ายตัว s ดังรูป



$0<P(x)<1$

สมการของ Binary Logistic Regression คือ

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 X$$

เราจะเรียก $\log\left(\frac{p(X)}{1-p(X)}\right)$ ว่า logit หรือฟังก์ชั่น log-odds ก็ได้

LOGISTIC REGRESSION

และจะเรียก $\frac{p(X)}{1 - p(X)}$ ว่า odds ซึ่งเป็นอัตราส่วนของความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นได้สำเร็จเป็นกี่เท่าของความน่าจะเป็นที่เหตุการณ์จะเกิดขึ้นไม่สำเร็จ

จากสมการ ค่า $\beta 1$ หมายถึง การเปลี่ยนแปลงของเส้นโค้งเมื่อค่า x เพิ่มขึ้น

- ถ้า $\beta 1 > 0$ แล้ว เมื่อค่า x เพิ่มขึ้น จะทำให้ความน่าจะเป็นเพิ่มขึ้น
- ถ้า $\beta 1 < 0$ แล้ว เมื่อค่า x เพิ่มขึ้น จะทำให้ความน่าจะเป็นน้อยลง

ถ้าเรากลับค่าสมการ Binary Logistic Regression ด้านบน เราจะได้สมการที่ชื่อว่า Sigmoid function ออกมา สมการนี้ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจ ซึ่งก็คือค่า p(X) ดังนี้

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

ค่า p(X) จะมีค่ามากกว่า 0 แต่น้อย 1 เสมอ ยิ่งค่า p(X) เข้าใกล้ 1 มาก แปลว่า ความน่าจะเป็นที่จะเกิดเหตุการณ์นั้นสูง แต่ถ้าค่า p(X) เข้าใกล้ 0 มาก แปลว่า ความน่าจะเป็นที่จะเกิดเหตุการณ์นั้นต่ำ

LOGISTIC REGRESSION

หากเรากำหนดให้ q(X) คือ ความน่าจะเป็นที่จะไม่เกิดเหตุการณ์ที่สนใจแล้ว จะได้ว่า

$$q(X) = 1 - p(X) = 1 - \frac{e^{\beta_0 + \beta X}}{1 + e^{\beta_0 + \beta X}} = \frac{1}{1 + e^{\beta_0 + \beta X}}$$

ทั้งนี้หากจะวิเคราะห์การถดถอยโลจิสติกแบบ Multinomial Logistic Regression ให้ใช้สมการดังนี้

$$p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n}}$$

สูตรสมการที่กล่าวมานั้นเป็นหลักการเบื้องต้นของ Logistic Regression ที่ควรทราบ ซึ่งไลบรารี Scikit-learn ได้นำทฤษฎีดังกล่าวมาจัดทำเป็นอัลกอริทึม Logistic Regression ให้เรียกใช้งาน เพื่อให้เขียนโปรแกรมได้สะดวกและง่ายขึ้น ซึ่งนอกจากอัลกอริทึมนี้จะสามารถทำนายค่าข้อมูลได้แล้ว ยังสามารถตรวจสอบความถูกต้อง (accuracy) ของการทำงานของอัลกอริทึมได้ด้วย

LOGISTIC REGRESSION

ทำความรู้จักกับ Training data และ Testing data

โดยปกติเมื่อเราทำการโหลดข้อมูล Dataset เข้ามาในโปรแกรมสำหรับใช้สอน (Train) ให้กับคอมพิวเตอร์ เราจะไม่นำข้อมูลทั้งหมดมาใช้สอน แต่เราจะแบ่งข้อมูล (Split) ออกเป็น 2 ส่วน คือ

- **ส่วนของ Training data** หมายถึง ข้อมูลที่นำมาใช้สำหรับสอน (Train) ให้กับคอมพิวเตอร์ เพื่อให้คอมพิวเตอร์เรียนรู้และสร้างโมเดลการเรียนรู้ขึ้นมา
- **ส่วนของ Testing data** หมายถึง ข้อมูลที่นำมาป้อนให้กับคอมพิวเตอร์ เพื่อทดสอบว่าโมเดลการเรียนรู้ของคอมพิวเตอร์ที่สร้างขึ้นมานั้นมีประสิทธิภาพในการทำนายมากน้อยเพียงใด

Note

Dataset คือ ชุดของข้อมูลที่เก็บรวบรวมไว้สำหรับนำมาใช้สร้างโมเดลการเรียนรู้

ตัวอย่าง : วิเคราะห์แนวโน้มการเป็นมะเร็งเต้านม

LOGISTIC REGRESSION

Confusion matrix and classification report

จากฟังก์ชั่น accuracy_score() ที่ใช้ในการเปรียบเทียบเปอร์เซ็นต์ความถูกต้องของผลลัพธ์จริงกับผลลัพธ์ของโมเดลที่สร้างขึ้น เราสามารถนำมาแจกแจงเป็นตัวเลขในรูปแบบเมตริกซ์เรียกว่า Confusion Matrix ได้ดังรูป

ค่าจริงของข้อมูล	ค่าที่โมเดลทำนาย: Positive	ค่าที่โมเดลทำนาย: Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

ภายใน Confusion Matrix ประกอบด้วยค่าดังต่อไปนี้

- **True Positive (TP)** ข้อมูลมีค่าผลลัพธ์เป็น "จริง" และโมเดลทำนายว่าเป็น "จริง" เหมือนกัน
- **True Negative (TN)** ข้อมูลมีค่าผลลัพธ์เป็น "เท็จ" และโมเดลทำนายว่าเป็น "เท็จ" เหมือนกัน
- **False Positive (FP)** ข้อมูลมีค่าผลลัพธ์เป็น "เท็จ" แต่โมเดลทำนายว่าเป็น "จริง"
- **False Negative (FN)** ข้อมูลมีค่าผลลัพธ์เป็น "จริง" แต่โมเดลทำนายว่าเป็น "เท็จ"

LOGISTIC REGRESSION

Confusion matrix and classification report

จากตัวอย่างเมื่อเรามีการเรียกใช้ฟังก์ชั่น accuracy_score() เพื่อเปรียบเทียบเลเบลของผลลัพธ์จริงกับเลเบลของผลลัพธ์ที่ได้จากการทำนาย จึงทำให้เกิดการคำนวณค่า Accuracy จาก Confusion Matrix ดังสูตรต่อไปนี้

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

จากสูตร นั่นคือ ค่า Accuracy จะถูกคำนวณจากผลลัพธ์ที่ทำนายได้ถูกต้องตรงกับเลเบลของผลลัพธ์จริงทุกกรณีเทียบกับผลลัพธ์ทั้งหมดนั่นเอง

LOGISTIC REGRESSION

Confusion matrix and classification report

ค่าที่สรุปใน Confustion Matrix จะสามารถนำมาวิเคราะห์ความแม่นยำของโมเดลได้ด้วยการคำนวณพารามิเตอร์เพื่อเป็นตัวชี้วัด 3 ตัวที่สำคัญดังนี้

- **Precision**

คือ จำนวนครั้งที่โมเดลทำนายว่าเป็นจริงแล้วทำนายถูก เป็นอัตราส่วนเท่าใดของ จำนวนครั้งทั้งหมดที่โมเดลทำนายว่าเป็นจริง

$$Precision = \frac{TP}{TP+FP}$$

- **Recall หรือ Sensitivity**

คือ ค่าที่บอกว่าจำนวนครั้งที่โมเดลทำนายว่าเป็นจริงแล้วทำนายถูก เป็นอัตราส่วนเท่าใดของจำนวนครั้งทั้งหมดที่ผลลัพธ์เป็นจริง

$$Recall = \frac{TP}{TP+FN}$$

- **F1-Score**

เป็นค่าเฉลี่ยกลางแบบ Weight Average ระหว่าง Precision และ Recall เพื่อใช้บอกความแม่นยำโดยรวมของการทำนายผลลัพธ์นั้น โดยไม่ต้องสนใจค่าของ Precision หรือ Recall โดยมีสูตรการคำนวณดังนี้

$$F1\text{-}Score = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

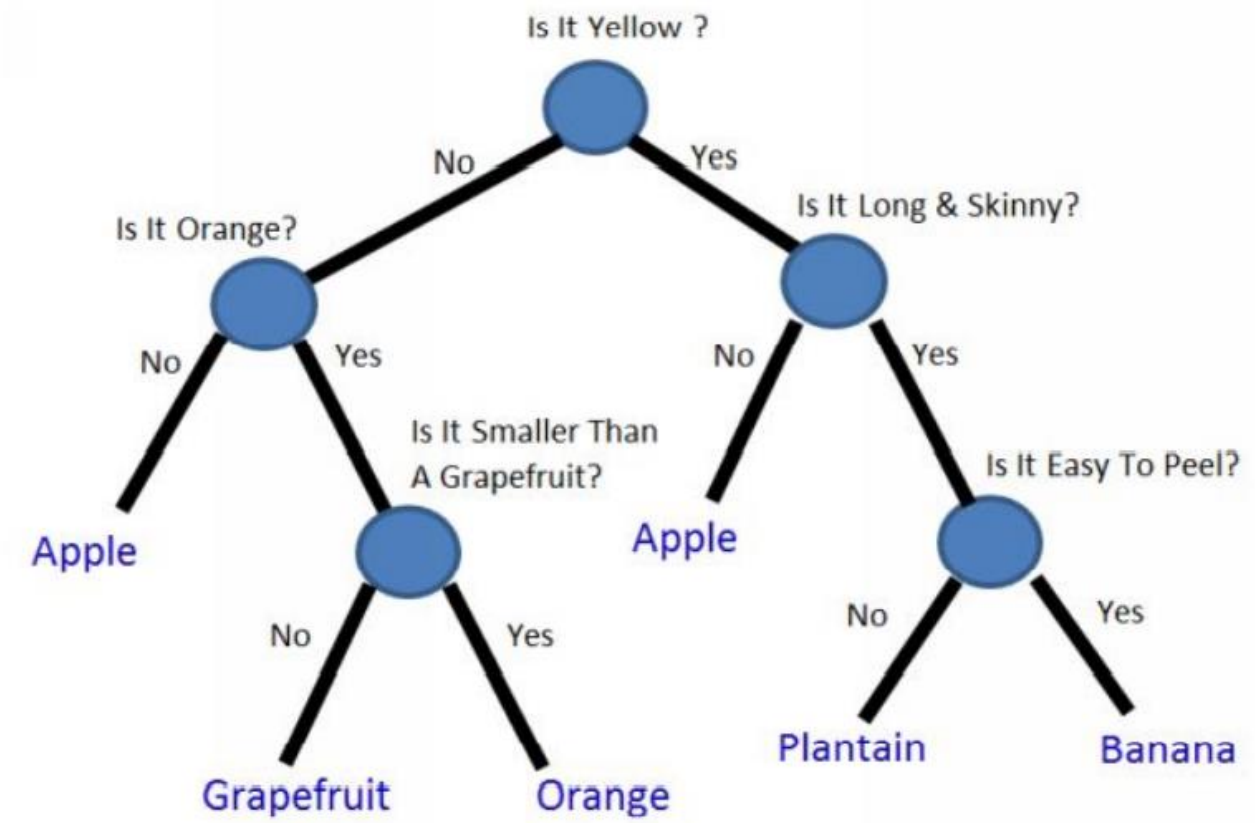
ML ALGORITHM SUMMARY

Start With a Data Set

DECISION TREE ALGORITHM

- A Decision Tree is simply a step by step process to go through to decide a category something belongs to.
- For example, let's say that you had a basket of fruit in front of you, and you were trying to teach someone who had never seen these types of fruit before how to tell them apart. How could you do it?

- Is it yellow?
- If so, is it long and skinny?
- If so, is it easy to peel?
- Then it is a banana



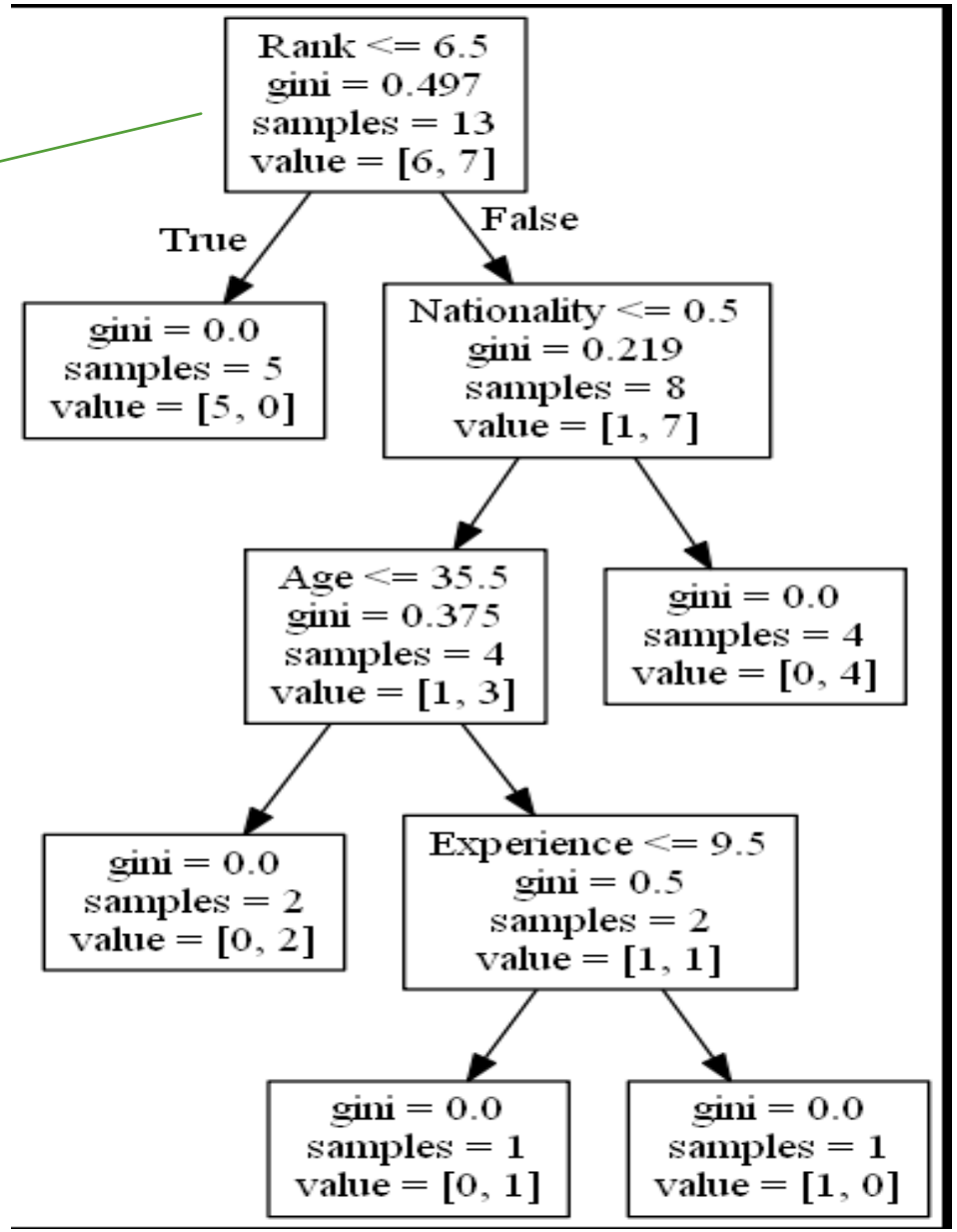
DECISION TREE ALGORITHM

Example: a person will try to decide if he/she should go to a comedy show or not.

Person has registered every time there was a comedy show in town, and registered some information about the comedian, and also registered if he/she went or not.

Age	Experience	Rank	Nationality	Go
36	10	9	UK	NO
42	12	4	USA	NO
23	4	6	N	NO
52	4	4	USA	NO
43	21	8	USA	YES
44	14	5	UK	NO
66	3	7	N	YES
35	14	9	UK	YES
52	13	7	N	YES
35	5	9	N	YES
24	3	5	USA	NO
18	3	7	UK	YES
45	9	9	UK	YES

DECISION TREE



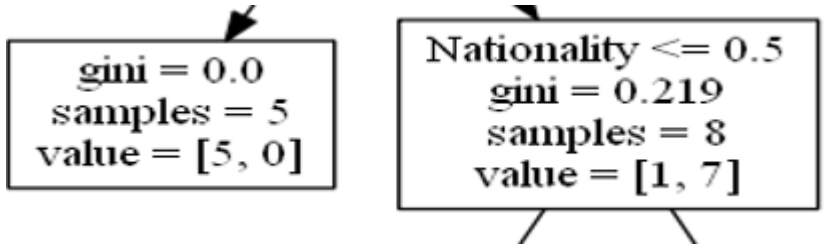
Rank
Rank <= 6.5 means that every comedian with a rank of 6.5 or lower will follow the True arrow (to the left), and the rest will follow the False arrow (to the right).

gini = 0.497 refers to the quality of the split, and is always a number between 0.0 and 0.5, where 0.0 would mean all of the samples got the same result, and 0.5 would mean that the split is done exactly in the middle.

samples = 13 means that there are 13 comedians left at this point in the decision, which is all of them since this is the first step.

value = [6, 7] means that of these 13 comedians, 6 will get a "NO", and 7 will get a "GO".

DECISION TREE



The next step contains two boxes, one box for the comedians with a 'Rank' of 6.5 or lower, and one box with the rest.

True - 5 Comedians End Here:

`gini = 0.0` means all of the samples got the same result.

`samples = 5` means that there are 5 comedians left in this branch (5 comedian with a Rank of 6.5 or lower).

`value = [5, 0]` means that 5 will get a "NO" and 0 will get a "GO".

False - 8 Comedians Continue:

Gini

There are many ways to split the samples, we use the GINI method in this tutorial.

The Gini method uses this formula:

$Gini = 1 - (x/n)^2 - (y/n)^2$

Where `x` is the number of positive answers("GO"), `n` is the number of samples, and `y` is the number of negative answers ("NO"), which gives us this calculation:

$1 - (7 / 13)^2 - (6 / 13)^2 = 0.497$

Nationality

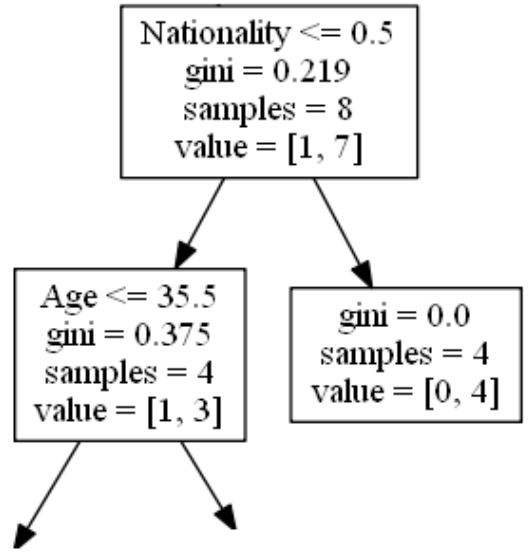
`Nationality <= 0.5` means that the comedians with a nationality value of less than 0.5 will follow the arrow to the left (which means everyone from the UK,), and the rest will follow the arrow to the right.

`gini = 0.219` means that about 22% of the samples would go in one direction.

`samples = 8` means that there are 8 comedians left in this branch (8 comedian with a Rank higher than 6.5).

`value = [1, 7]` means that of these 8 comedians, 1 will get a "NO" and 7 will get a "GO".

DECISION TREE



True - 4 Comedians Continue:

Age

`Age <= 35.5` means that comedians at the age of 35.5 or younger will follow the arrow to the left, and the rest will follow the arrow to the right.

`gini = 0.375` means that about 37,5% of the samples would go in one direction.

`samples = 4` means that there are 4 comedians left in this branch (4 comedians from the UK).

`value = [1, 3]` means that of these 4 comedians, 1 will get a "NO" and 3 will get a "GO".

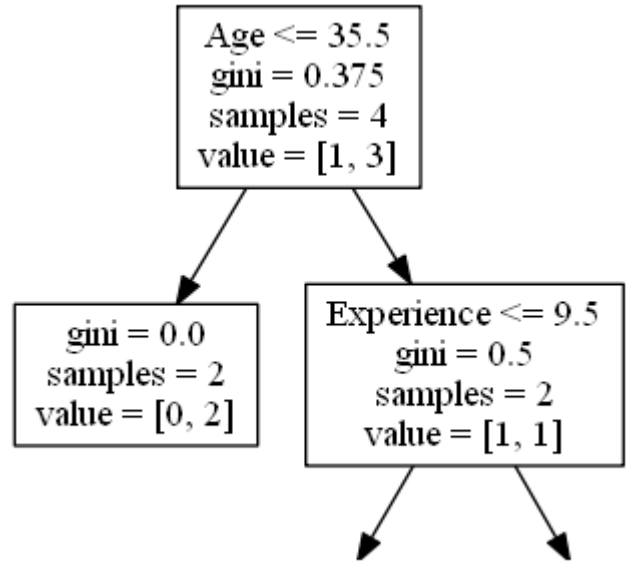
False - 4 Comedians End Here:

`gini = 0.0` means all of the samples got the same result.

`samples = 4` means that there are 4 comedians left in this branch (4 comedians not from the UK).

`value = [0, 4]` means that of these 4 comedians, 0 will get a "NO" and 4 will get a "GO".

DECISION TREE



True - 2 Comedians End Here:

`gini = 0.0` means all of the samples got the same result.

`samples = 2` means that there are 2 comedians left in this branch (2 comedians at the age 35.5 or younger).

`value = [0, 2]` means that of these 2 comedians, 0 will get a "NO" and 2 will get a "GO".

False - 2 Comedians Continue:

False - 2 Comedians Continue:

Experience

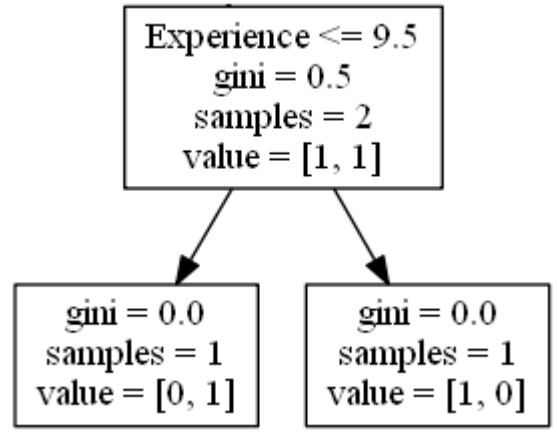
`Experience <= 9.5` means that comedians with 9.5 years of experience, or less, will follow the arrow to the left, and the rest will follow the arrow to the right.

`gini = 0.5` means that 50% of the samples would go in one direction.

`samples = 2` means that there are 2 comedians left in this branch (2 comedians older than 35.5).

`value = [1, 1]` means that of these 2 comedians, 1 will get a "NO" and 1 will get a "GO".

DECISION TREE



True - 1 Comedian Ends Here:

`gini = 0.0` means all of the samples got the same result.

`samples = 1` means that there is 1 comedian left in this branch (1 comedian with 9.5 years of experience or less).

`value = [0, 1]` means that 0 will get a "NO" and 1 will get a "GO".

False - 1 Comedian Ends Here:

`gini = 0.0` means all of the samples got the same result.

`samples = 1` means that there is 1 comedians left in this branch (1 comedian with more than 9.5 years of experience).

`value = [1, 0]` means that 1 will get a "NO" and 0 will get a "GO".

DECISION TREE

Predict Values

We can use the Decision Tree to predict new values.

Example: Should I go see a show starring a 40 years old American comedian, with 10 years of experience, and a comedy ranking of 7?

Example

Use predict() method to predict new values:

```
print(dtree.predict([[40, 10, 7, 1]]))
```

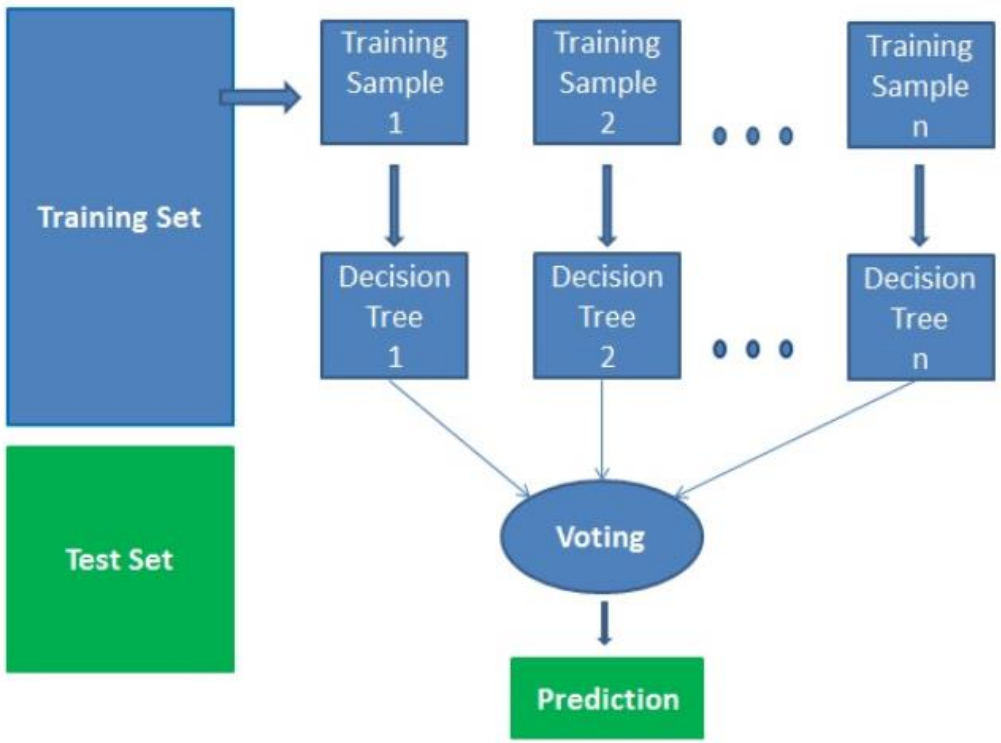
[0]
[1] means 'GO'
[0] means 'NO'

RANDOM FOREST

How does the algorithm work?

It works in four steps:

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.



RANDOM FOREST

Advantages:

- Random forests is considered as a highly accurate and robust method because of the number of decision trees participating in the process.
- It does not suffer from the overfitting problem. The main reason is that it takes the average of all the predictions, which cancels out the biases.
- The algorithm can be used in both classification and regression problems.
- Random forests can also handle missing values. There are two ways to handle these: using median values to replace continuous variables, and computing the proximity-weighted average of missing values.
- You can get the relative feature importance, which helps in selecting the most contributing features for the classifier.

Disadvantages:

- Random forests is slow in generating predictions because it has multiple decision trees. Whenever it makes a prediction, all the trees in the forest have to make a prediction for the same given input and then perform voting on it. This whole process is time-consuming.
- The model is difficult to interpret compared to a decision tree, where you can easily make a decision by following the path in the tree.

RANDOM FOREST

Iris versicolor

Iris virginica

Classification

HOMEWORK

1. Logistic Regression

Data

The dataset comes from the UCI Machine Learning repository, and it is related to direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict whether the client will subscribe (1/0) to a term deposit (variable y). The dataset can be downloaded from here.

https://towardsdatascience.com/building-a-logistic-regression-in-python-step-by-step-becd4d56c9c8

2. Random Forest – temperature prediction temperature for tomorrow

https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

Problem Introduction

The problem we will tackle is predicting the max temperature for tomorrow in our city using one year of past weather data. I am using Seattle, WA but feel free to find data for your own city using the NOAA Climate Data Online tool. We are going to act as if we don't have access to any weather forecasts (and besides, it's more fun to make our own predictions rather than rely on others). What we do have access to is one year of historical max temperatures, the temperatures for the previous two days, and an estimate

Data Acquisition

s available for download :

HOMEWORK

2. Random Forest – temperature prediction

THE END