Introduction to Data Science

Last Update: 24 JAN 2020

Chapter 4
Predictive Analysis

Papangkorn Inkeaw, PhD

Department of Computer Science, Faculty of Science
Chiang Mai University

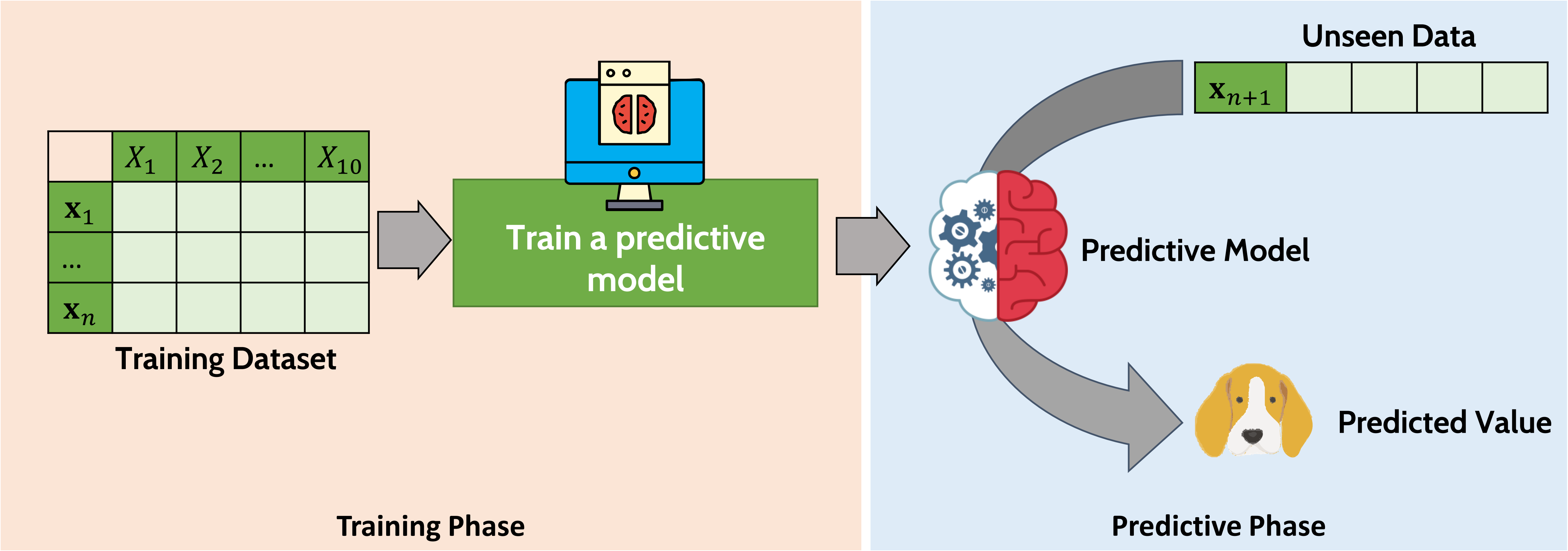
Outline

Predictive Analysis

1. Predictive Analysis
   - Preparing Datasets
2. Classification Analysis
   - K-Nearest Neighbor
   - Decision Tree
   - Naïve Bayes
   - Classification Assessment
3. Regression Analysis
   - Linear Regression
   - Polynomial Regression
   - Regression Assessment
4. Time Series Analysis
   - Moving Average
   - Autoregressive Integrated Moving
   - Curve Fitting

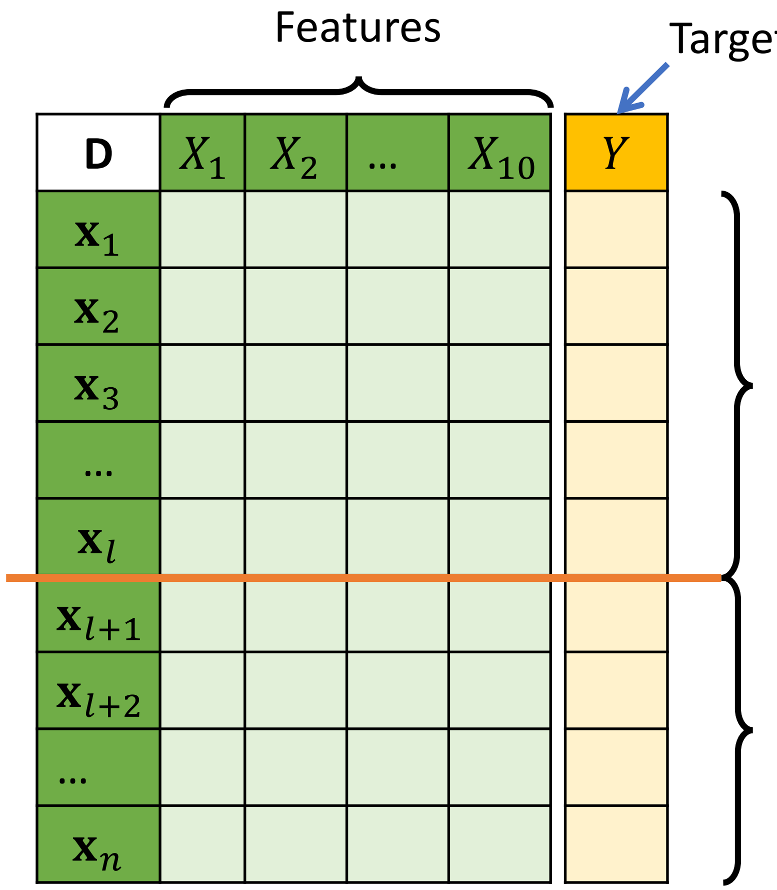
Predictive Analysis

Analyze current and historical data to make predictions about future or otherwise unknown events.



Preparing Dataset

Predictive Analysis



To perform a predictive analysis:

- We should have two dataset: training and test datasets.
- The target value of each datapoint must be available.

Training dataset

- Will be used to train a predictive model.
- Target value of each data point must be available.

Test dataset

- Will be used to evaluate the predictive model
- Assume that target value of each data point is not known, but it should be available.

Classification Analysis

Features

Target class

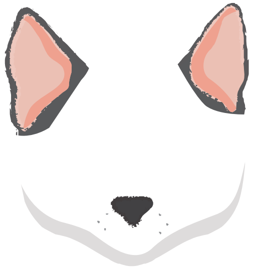
D	X_1	X_2	…	X_{10}	Y
$\mathbf{x}_1$					
$\mathbf{x}_2$					
$\mathbf{x}_3$					
…					
$\mathbf{x}_l$					
$\mathbf{x}_{l+1}$					
$\mathbf{x}_{l+2}$					
…					
$\mathbf{x}_n$					

For classification analysis

- The value we want to predict is **categorical data.**
- Known as **class**

Example

We know some characteristics of an animal, and we want to predict it is a cat or a dog.



cat or dog?

Classification Analysis

Types of Classification Problems

Number of Possible Classes

= 2

> 2

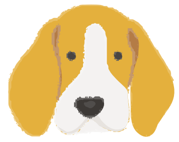
Binary Classification

- Want to distinguish only two categories (classes)

Cat or Dog?

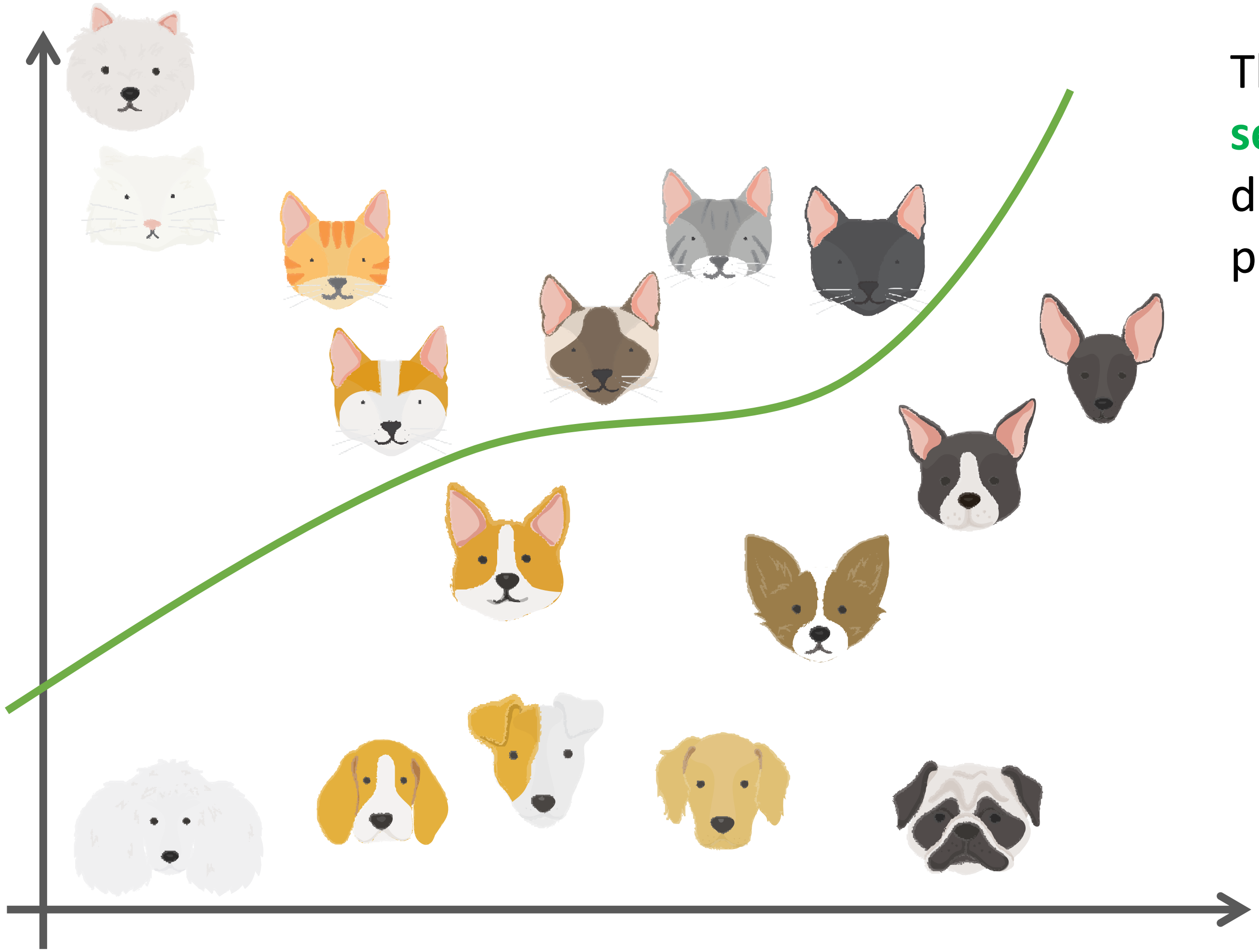
Multi-class Classification

- Want to distinguish more than two categories (classes)



Bulldog or Beagle or Shiba ?

Classification Analysis



The task of classification is one of finding **separating lines** that separate classes of data from a training dataset as best as possible.

K-Nearest Neighbor

Classification Analysis

K-Nearest Neighbor classifier assigns the class label of an unseen data with the majority class labels of k neighbor data (in the training dataset)

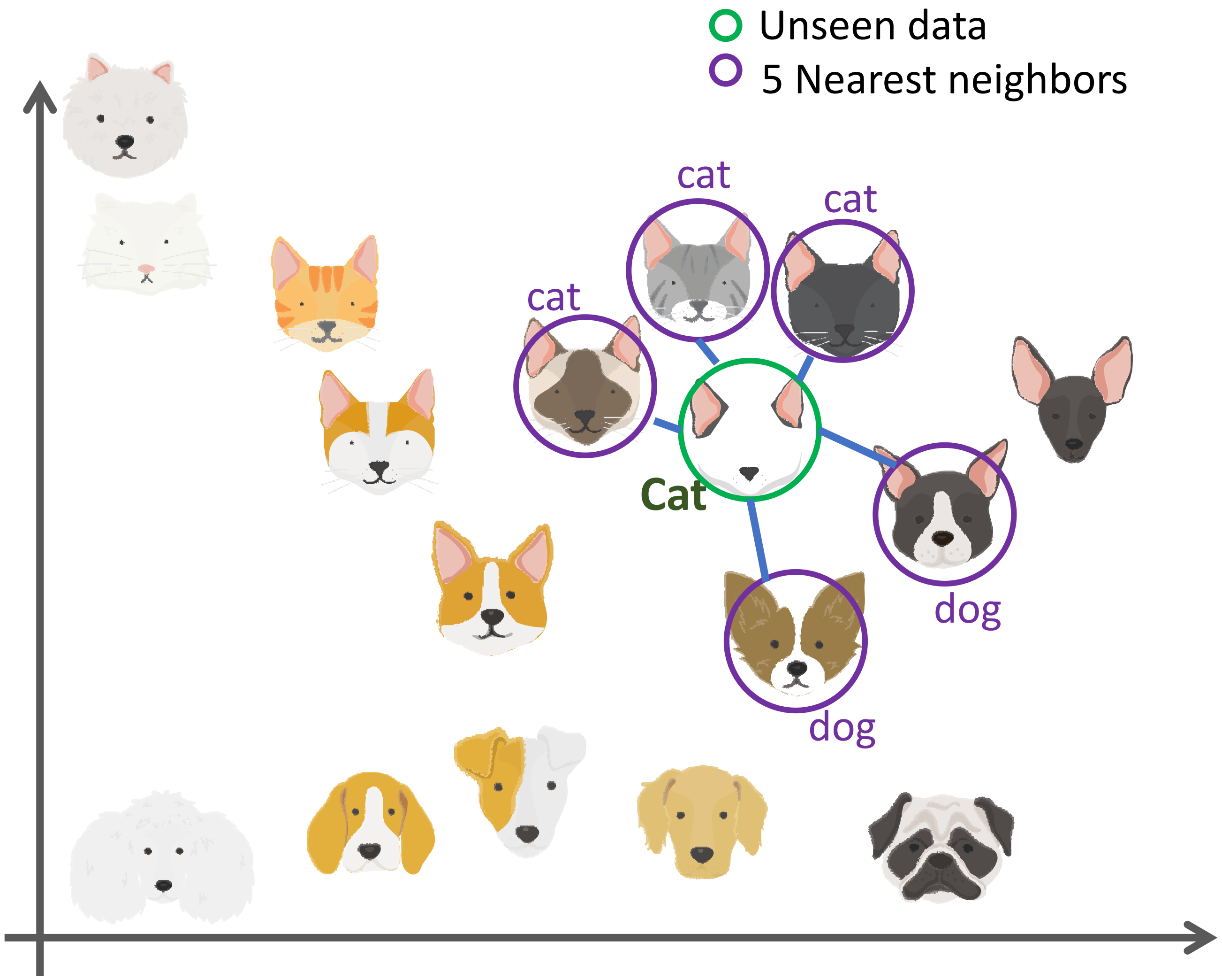
How the k-nearest neighbor works

STEP 1: Calculate distances between an unseen data and training data

STEP 2: Find *k* nearest neighbor

STEP 3: Find majority class label

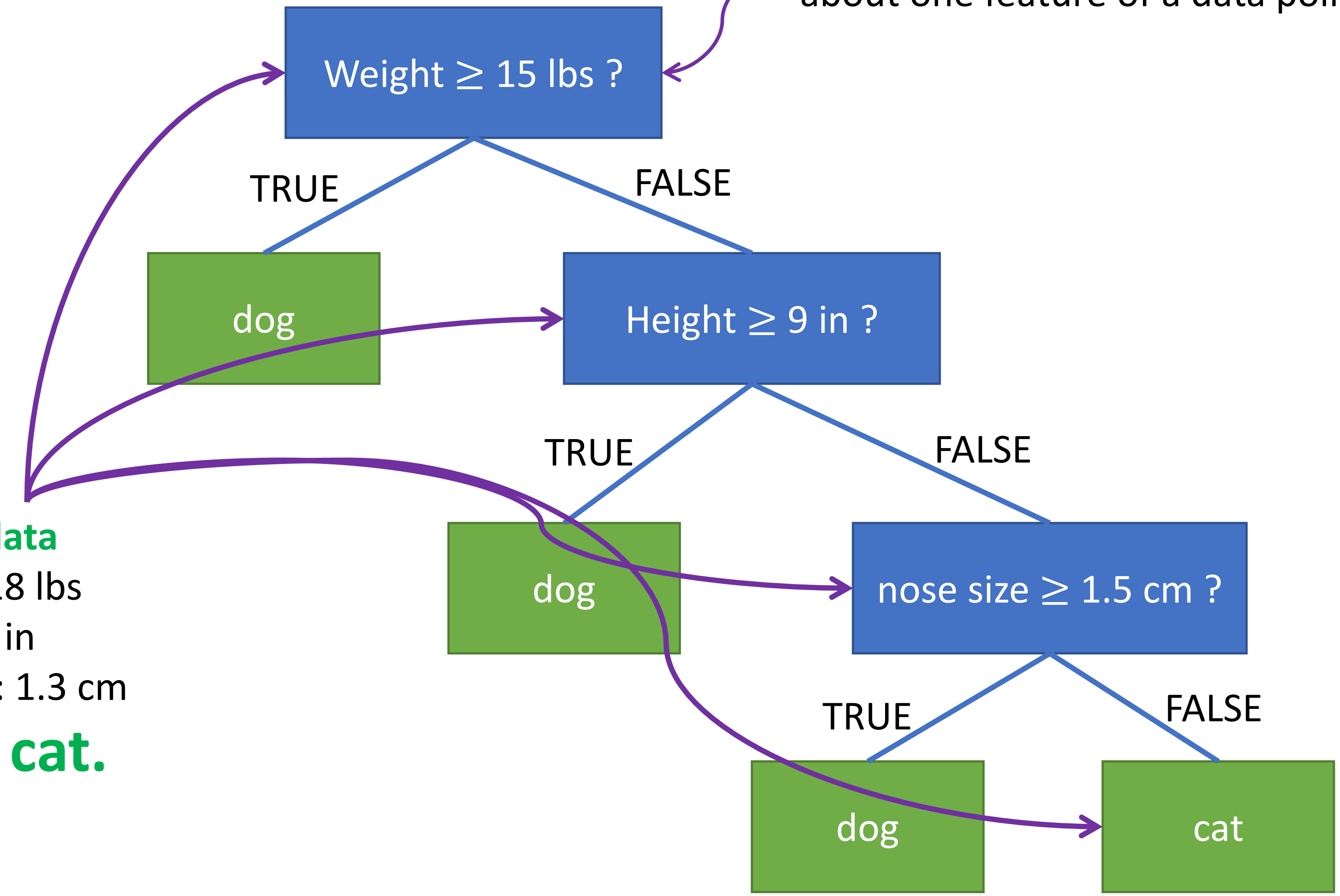
STEP 4: Assign the majority class label to the class label of the unseen data



Decision Tree

Classification Analysis

Every node in the tree asks a question about one feature of a data point.



Unseen data
Weight: 18 lbs
Height: 9 in
nose size: 1.3 cm

So, it is a cat.

Decision Tree

Classification Analysis

Construct a decision tree

STEP 1: Given a training data D, find the single feature (and cutoff for that feature, if it's numerical) that best partitions your data into classes.

STEP 2: This single best feature/cutoff becomes the root of your decision tree.

STEP 3: Partition *D* up according to the root node.

STEP 4: Recursively train each of the child nodes on its partition of the data until all of the data points in the partition have the same label.

D	Weight	Height	Nose size	Label
$\mathbf{x}_1$	8	8	1.6	Dog
$\mathbf{x}_2$	50	40	3	Dog
$\mathbf{x}_3$	8	9	1.3	Cat
$\mathbf{x}_4$	15	12	2.5	Dog
$\mathbf{x}_5$	9	9.8	1.4	Cat

Weight $\geq$ 15 lbs ?

TRUE

D	Weight	Height	Nose size	Label
$\mathbf{x}_2$	50	40	3	Dog
$\mathbf{x}_4$	15	12	2.5	Dog

FALSE

D	Weight	Height	Nose size	Label
$\mathbf{x}_1$	8	8	1.6	Dog
$\mathbf{x}_3$	8	9	1.3	Cat
$\mathbf{x}_5$	9	9.8	1.4	Cat

Decision Tree

Classification Analysis

Construct a decision tree

STEP 1: Given a training data D, find the single feature (and cutoff for that feature, if it's numerical) that best partitions your data into classes.
STEP 2: This single best feature/cutoff becomes the root of your decision tree.
STEP 3: Partition *D* up according to the root node.
STEP 4: Recursively train each of the child nodes on its partition of the data until all of the data points in the partition have the same label.

Weight $\geq$ 15 lbs ?

FALSE

D	Weight	Height	Nose size	Label
$\mathbf{x}_1$	8	8	1.6	Dog
$\mathbf{x}_3$	8	9	1.3	Cat
$\mathbf{x}_5$	9	8.5	1.4	Cat

Height $\geq$ 9 in ?

TRUE

D	Weight	Height	Nose size	Label
$\mathbf{x}_3$	8	9	1.3	Cat

FALSE

D	Weight	Height	Nose size	Label
$\mathbf{x}_1$	8	8	1.6	Dog
$\mathbf{x}_5$	9	8.5	1.4	Cat

Decision Tree

Classification Analysis

Construct a decision tree

STEP 1: Given a training data D, find the single feature (and cutoff for that feature, if it's numerical) that best partitions your data into classes.

STEP 2: This single best feature/cutoff becomes the root of your decision tree.

STEP 3: Partition *D* up according to the root node.

STEP 4: Recursively train each of the child nodes on its partition of the data until all of the data points in the partition have the same label.

Height $\geq$ 9 in ?

FALSE

D	Weight	Height	Nose size	Label
$\mathbf{x}_1$	8	8	1.6	Dog
$\mathbf{x}_5$	9	8.5	1.4	Cat

nose size $\geq$ 1.5 cm ?

TRUE

D	Weight	Height	Nose size	Label
$\mathbf{x}_1$	8	8	1.6	Dog

FALSE

D	Weight	Height	Nose size	Label
$\mathbf{x}_5$	9	8.5	1.4	Cat

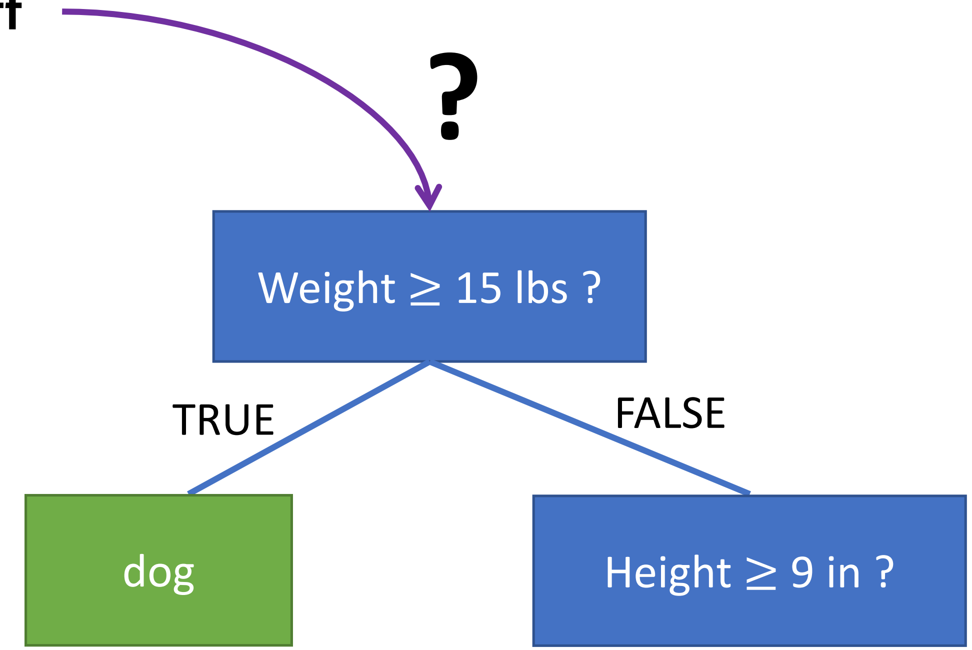
Decision Tree

Classification Analysis

How to determine the best feature and cutoff

The most common ones are:

- Information gain
- Gini impurity.



You can find more details in:

- Zaki, M., & Meira, W. (2014). Data mining and analysis : Fundamental concepts and algorithms. New York: Cambridge University Press.
- https://en.wikipedia.org/wiki/Decision_tree_learning

Naïve Bayes

Classification Analysis

Bayes Theorem:

The *prior*, the initial degree of belief in ***A***.

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Probability of ***A*** happening, given that ***B*** has occurred

The likelihood of event ***B*** occurring given that ***A*** is true.

Thomas Bayes
1701-1761

Source: https://en.wikipedia.org/wiki/Thomas_Bayes#/media/File:Thomas_Bayes.gif

Naïve Bayes

Classification Analysis

Classify whether the day is suitable for playing golf, given the features of the day.

Bayes theorem can be rewritten as:

$$P(y|\mathbf{x}) = \frac{P(y)P(\mathbf{x}|y)}{P(\mathbf{x})}$$

We want to classify

$$\mathbf{x} = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$$

D	Outlook	Temperature	Humidity	Windy	Play golf
$\mathbf{x}_1$	Rainy	Hot	High	False	No
$\mathbf{x}_2$	Rainy	Hot	High	True	No
$\mathbf{x}_3$	Overcast	Hot	High	False	Yes
$\mathbf{x}_4$	Sunny	Mild	High	False	Yes
$\mathbf{x}_5$	Sunny	Cool	Normal	False	Yes
$\mathbf{x}_6$	Sunny	Cool	Normal	True	No
$\mathbf{x}_7$	Overcast	Cool	Normal	True	Yes
$\mathbf{x}_8$	Rainy	Mild	High	False	No
$\mathbf{x}_9$	Rainy	Cool	Normal	False	Yes
$\mathbf{x}_{10}$	Sunny	Mild	Normal	False	Yes
$\mathbf{x}_{11}$	Rainy	Mild	Normal	True	Yes
$\mathbf{x}_{12}$	Overcast	Mild	High	Ture	Yes
$\mathbf{x}_{13}$	Overcast	Hot	Normal	False	Yes
$\mathbf{x}_{14}$	Sunny	Mild	High	True	No

Naïve Bayes

Classification Analysis

How the Naïve Bayes works

STEP 1: Calculate $P(y)$ for all possible value of y from the training dataset.

STEP 2: Calculate $P(\mathbf{x}|y) = \prod_{i=1}^{p} P(x_i|y)$ for all possible value of y from the training dataset.

STEP 3: Calculate $P(y|\mathbf{x}) = P(y) \prod_{i=1}^{p} P(x_i|y)$

STEP 4: Assign y that reach the highest $P(y|\mathbf{x})$ to the class label of $\mathbf{x}$

We want to classify

$$\mathbf{x} = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$$

D	Outlook	Temperature	Humidity	Windy	Play golf
$\mathbf{x}_1$	Rainy	Hot	High	False	No
$\mathbf{x}_2$	Rainy	Hot	High	True	No
$\mathbf{x}_3$	Overcast	Hot	High	False	Yes
$\mathbf{x}_4$	Sunny	Mild	High	False	Yes
$\mathbf{x}_5$	Sunny	Cool	Normal	False	Yes
$\mathbf{x}_6$	Sunny	Cool	Normal	True	No
$\mathbf{x}_7$	Overcast	Cool	Normal	True	Yes
$\mathbf{x}_8$	Rainy	Mild	High	False	No
$\mathbf{x}_9$	Rainy	Cool	Normal	False	Yes
$\mathbf{x}_{10}$	Sunny	Mild	Normal	False	Yes
$\mathbf{x}_{11}$	Rainy	Mild	Normal	True	Yes
$\mathbf{x}_{12}$	Overcast	Mild	High	Ture	Yes
$\mathbf{x}_{13}$	Overcast	Hot	Normal	False	Yes
$\mathbf{x}_{14}$	Sunny	Mild	High	True	No

Naïve Bayes

Classification Analysis

How the Naïve Bayes works

STEP 1: Calculate $P(y)$ for all possible value of y from the training dataset.

STEP 2: Calculate $P(\mathbf{x}|y) = \prod_{i=1}^{p} P(x_i|y)$ for all possible value of y from the training dataset.

STEP 3: Calculate $P(y|\mathbf{x}) = P(y) \prod_{i=1}^{p} P(x_i|y)$

STEP 4: Assign y that reach the highest $P(y|\mathbf{x})$ to the class label of $\mathbf{x}$

$$P(\text{Play golf} = \text{No}) = \frac{5}{14}$$

$$P(\text{Play golf} = \text{Yes}) = \frac{9}{14}$$

We want to classify

$$\mathbf{x} = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$$

D	Outlook	Temperature	Humidity	Windy	Play golf
$\mathbf{x}_1$	Rainy	Hot	High	False	No
$\mathbf{x}_2$	Rainy	Hot	High	True	No
$\mathbf{x}_3$	Overcast	Hot	High	False	Yes
$\mathbf{x}_4$	Sunny	Mild	High	False	Yes
$\mathbf{x}_5$	Sunny	Cool	Normal	False	Yes
$\mathbf{x}_6$	Sunny	Cool	Normal	True	No
$\mathbf{x}_7$	Overcast	Cool	Normal	True	Yes
$\mathbf{x}_8$	Rainy	Mild	High	False	No
$\mathbf{x}_9$	Rainy	Cool	Normal	False	Yes
$\mathbf{x}_{10}$	Sunny	Mild	Normal	False	Yes
$\mathbf{x}_{11}$	Rainy	Mild	Normal	True	Yes
$\mathbf{x}_{12}$	Overcast	Mild	High	Ture	Yes
$\mathbf{x}_{13}$	Overcast	Hot	Normal	False	Yes
$\mathbf{x}_{14}$	Sunny	Mild	High	True	No

Naïve Bayes

Classification Analysis

How the Naïve Bayes works

STEP 1: Calculate $P(y)$ for all possible value of y from the training dataset.

STEP 2: Calculate $P(\mathbf{x}|y) = \prod_{i=1}^{p} P(x_i|y)$ for all possible value of y from the training dataset.

STEP 3: Calculate $P(y|\mathbf{x}) = P(y)\prod_{i=1}^{p} P(x_i|y)$

STEP 4: Assign y that reach the highest $P(y|\mathbf{x})$ to the class label of $\mathbf{x}$

$$P(\text{Outlook} = \text{Sunny}|\text{Play golf} = \text{No}) = \frac{2}{5}$$

$$P(\text{Outlook} = \text{Sunny}|\text{Play golf} = \text{Yes}) = \frac{3}{9}$$

We want to classify

$$\mathbf{x} = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$$

D	Outlook	Temperature	Humidity	Windy	Play golf
$\mathbf{x}_1$	Rainy	Hot	High	False	No
$\mathbf{x}_2$	Rainy	Hot	High	True	No
$\mathbf{x}_3$	Overcast	Hot	High	False	Yes
$\mathbf{x}_4$	Sunny	Mild	High	False	Yes
$\mathbf{x}_5$	Sunny	Cool	Normal	False	Yes
$\mathbf{x}_6$	Sunny	Cool	Normal	True	No
$\mathbf{x}_7$	Overcast	Cool	Normal	True	Yes
$\mathbf{x}_8$	Rainy	Mild	High	False	No
$\mathbf{x}_9$	Rainy	Cool	Normal	False	Yes
$\mathbf{x}_{10}$	Sunny	Mild	Normal	False	Yes
$\mathbf{x}_{11}$	Rainy	Mild	Normal	True	Yes
$\mathbf{x}_{12}$	Overcast	Mild	High	Ture	Yes
$\mathbf{x}_{13}$	Overcast	Hot	Normal	False	Yes
$\mathbf{x}_{14}$	Sunny	Mild	High	True	No

Naïve Bayes

Classification Analysis

How the Naïve Bayes works

STEP 1: Calculate $P(y)$ for all possible value of y from the training dataset.

STEP 2: Calculate $P(\mathbf{x}|y) = \prod_{i=1}^{p} P(x_i|y)$ for all possible value of y from the training dataset.

STEP 3: Calculate $P(y|\mathbf{x}) = P(y) \prod_{i=1}^{p} P(x_i|y)$

STEP 4: Assign y that reach the highest $P(y|\mathbf{x})$ to the class label of $\mathbf{x}$

$$P(\text{Temperature} = \text{Hot}|\text{Play golf} = \text{No}) = \frac{2}{5}$$

$$P(\text{Temperature} = \text{Hot}|\text{Play golf} = \text{Yes}) = \frac{2}{9}$$

We want to classify

$$\mathbf{x} = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$$

D	Outlook	Temperature	Humidity	Windy	Play golf
$\mathbf{x}_1$	Rainy	Hot	High	False	No
$\mathbf{x}_2$	Rainy	Hot	High	True	No
$\mathbf{x}_3$	Overcast	Hot	High	False	Yes
$\mathbf{x}_4$	Sunny	Mild	High	False	Yes
$\mathbf{x}_5$	Sunny	Cool	Normal	False	Yes
$\mathbf{x}_6$	Sunny	Cool	Normal	True	No
$\mathbf{x}_7$	Overcast	Cool	Normal	True	Yes
$\mathbf{x}_8$	Rainy	Mild	High	False	No
$\mathbf{x}_9$	Rainy	Cool	Normal	False	Yes
$\mathbf{x}_{10}$	Sunny	Mild	Normal	False	Yes
$\mathbf{x}_{11}$	Rainy	Mild	Normal	True	Yes
$\mathbf{x}_{12}$	Overcast	Mild	High	Ture	Yes
$\mathbf{x}_{13}$	Overcast	Hot	Normal	False	Yes
$\mathbf{x}_{14}$	Sunny	Mild	High	True	No

Naïve Bayes

Classification Analysis

How the Naïve Bayes works

STEP 1: Calculate $P(y)$ for all possible value of y from the training dataset.

STEP 2: Calculate $P(\mathbf{x}|y) = \prod_{i=1}^{p} P(x_i|y)$ for all possible value of y from the training dataset.

STEP 3: Calculate $P(y|\mathbf{x}) = P(y)\prod_{i=1}^{p} P(x_i|y)$

STEP 4: Assign y that reach the highest $P(y|\mathbf{x})$ to the class label of $\mathbf{x}$

$$P(\text{Humidity} = \text{Normal}|\text{Play golf} = \text{No}) = \frac{1}{5}$$

$$P(\text{Humidity} = \text{Normal}|\text{Play golf} = \text{Yes}) = \frac{6}{9}$$

We want to classify

$$\mathbf{x} = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$$

D	Outlook	Temperature	Humidity	Windy	Play golf
$\mathbf{x}_1$	Rainy	Hot	High	False	No
$\mathbf{x}_2$	Rainy	Hot	High	True	No
$\mathbf{x}_3$	Overcast	Hot	High	False	Yes
$\mathbf{x}_4$	Sunny	Mild	High	False	Yes
$\mathbf{x}_5$	Sunny	Cool	Normal	False	Yes
$\mathbf{x}_6$	Sunny	Cool	Normal	True	No
$\mathbf{x}_7$	Overcast	Cool	Normal	True	Yes
$\mathbf{x}_8$	Rainy	Mild	High	False	No
$\mathbf{x}_9$	Rainy	Cool	Normal	False	Yes
$\mathbf{x}_{10}$	Sunny	Mild	Normal	False	Yes
$\mathbf{x}_{11}$	Rainy	Mild	Normal	True	Yes
$\mathbf{x}_{12}$	Overcast	Mild	High	Ture	Yes
$\mathbf{x}_{13}$	Overcast	Hot	Normal	False	Yes
$\mathbf{x}_{14}$	Sunny	Mild	High	True	No

Naïve Bayes

Classification Analysis

How the Naïve Bayes works

STEP 1: Calculate $P(y)$ for all possible value of y from the training dataset.

STEP 2: Calculate $P(\mathbf{x}|y) = \prod_{i=1}^{p} P(x_i|y)$ for all possible value of y from the training dataset.

STEP 3: Calculate $P(y|\mathbf{x}) = P(y) \prod_{i=1}^{p} P(x_i|y)$

STEP 4: Assign y that reach the highest $P(y|\mathbf{x})$ to the class label of $\mathbf{x}$

$$P(\text{Windy} = \text{True}|\text{Play golf} = \text{No}) = \frac{3}{5}$$

$$P(\text{Windy} = \text{True}|\text{Play golf} = \text{Yes}) = \frac{3}{9}$$

We want to classify

$$\mathbf{x} = (\text{Sunny}, \text{Hot}, \text{Normal}, \text{True})$$

D	Outlook	Temperature	Humidity	Windy	Play golf
$\mathbf{x}_1$	Rainy	Hot	High	False	No
$\mathbf{x}_2$	Rainy	Hot	High	True	No
$\mathbf{x}_3$	Overcast	Hot	High	False	Yes
$\mathbf{x}_4$	Sunny	Mild	High	False	Yes
$\mathbf{x}_5$	Sunny	Cool	Normal	False	Yes
$\mathbf{x}_6$	Sunny	Cool	Normal	True	No
$\mathbf{x}_7$	Overcast	Cool	Normal	True	Yes
$\mathbf{x}_8$	Rainy	Mild	High	False	No
$\mathbf{x}_9$	Rainy	Cool	Normal	False	Yes
$\mathbf{x}_{10}$	Sunny	Mild	Normal	False	Yes
$\mathbf{x}_{11}$	Rainy	Mild	Normal	True	Yes
$\mathbf{x}_{12}$	Overcast	Mild	High	Ture	Yes
$\mathbf{x}_{13}$	Overcast	Hot	Normal	False	Yes
$\mathbf{x}_{14}$	Sunny	Mild	High	True	No

Naïve Bayes

Classification Analysis

How the Naïve Bayes works

STEP 1: Calculate $P(y)$ for all possible value of y from the training dataset.

STEP 2: Calculate $P(\mathbf{x}|y) = \prod_{i=1}^{p} P(x_i|y)$ for all possible value of y from the training dataset.

STEP 3: Calculate $P(y|\mathbf{x}) = P(y)\prod_{i=1}^{p} P(x_i|y)$

STEP 4: Assign y that reach the highest $P(y|\mathbf{x})$ to the class label of $\mathbf{x}$

$$P(\text{Play golf} = \text{No}|\text{Sunny, Hot, Normal, True})$$
$$= \frac{5}{14} \times \frac{2}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{3}{5} = 0.0069$$

$$P(\text{Play golf} = \text{Yes}|\text{Sunny, Hot, Normal, True})$$
$$= \frac{9}{14} \times \frac{3}{9} \times \frac{2}{9} \times \frac{6}{9} \times \frac{3}{9} = \mathbf{0.0106}$$

So, it is suitable to **play golf** given the conditions (Outlook = Sunny, Temperature = Hot, Humidity = Normal and Windy = True).

We want to classify

$$\mathbf{x} = (\text{Sunny, Hot, Normal, True})$$

$$P(\text{Play golf} = \text{No}) = \frac{5}{14}$$

$$P(\text{Play golf} = \text{Yes}) = \frac{9}{14}$$

$$P(\text{Outlook} = \text{Sunny}|\text{Play golf} = \text{No}) = \frac{2}{5}$$

$$P(\text{Outlook} = \text{Sunny}|\text{Play golf} = \text{Yes}) = \frac{3}{9}$$

$$P(\text{Temperature} = \text{Hot}|\text{Play golf} = \text{No}) = \frac{2}{5}$$

$$P(\text{Temperature} = \text{Hot}|\text{Play golf} = \text{Yes}) = \frac{2}{9}$$

$$P(\text{Humidity} = \text{Normal}|\text{Play golf} = \text{No}) = \frac{1}{5}$$

$$P(\text{Humidity} = \text{Normal}|\text{Play golf} = \text{Yes}) = \frac{6}{9}$$

$$P(\text{Windy} = \text{True}|\text{Play golf} = \text{No}) = \frac{3}{5}$$

$$P(\text{Windy} = \text{True}|\text{Play golf} = \text{Yes}) = \frac{3}{9}$$

Naïve Bayes

Classification Analysis

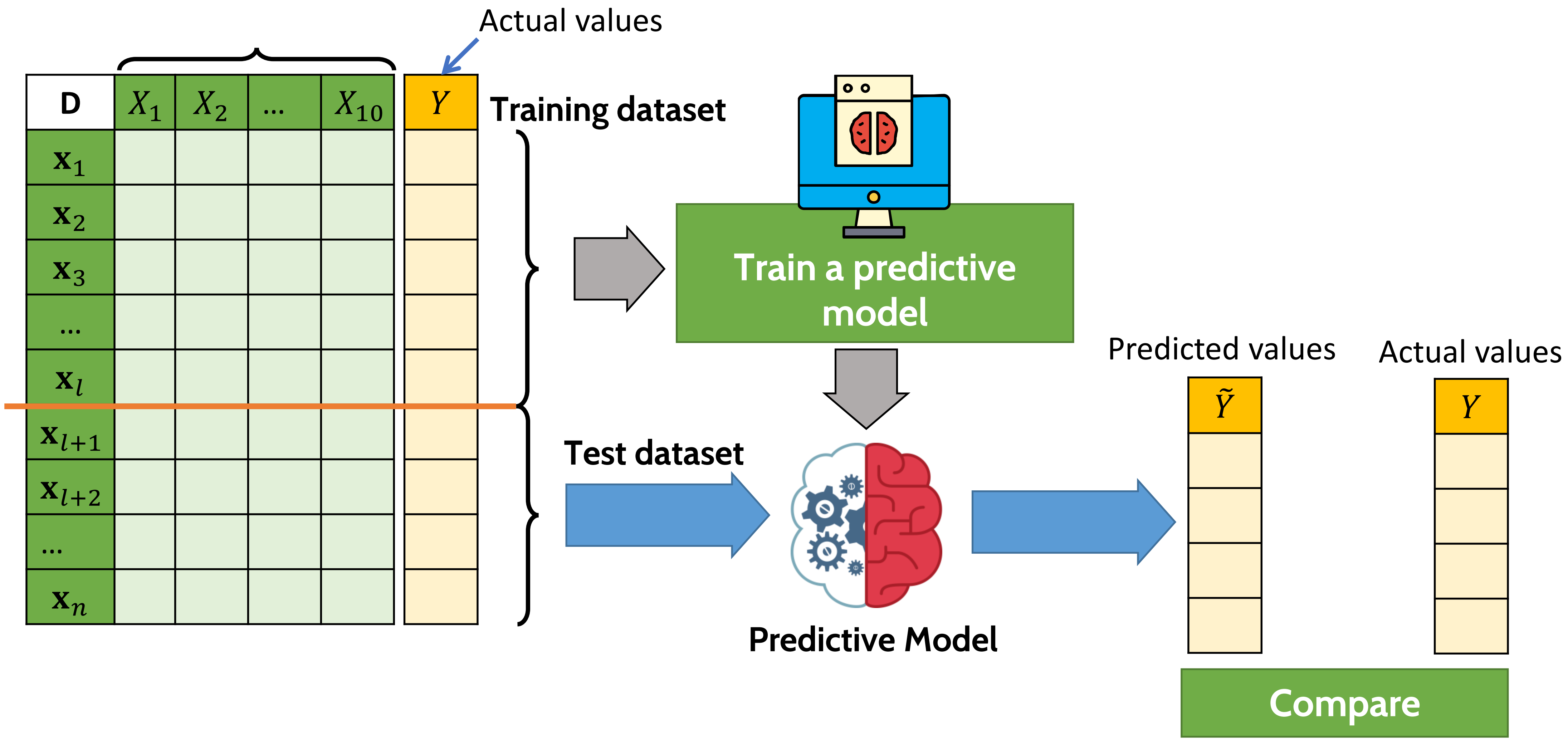
Quiz:

It is suitable to play golf or not given the conditions (Outlook = Rainy, Temperature = Mild, Humidity = Normal and Windy = False).

D	Outlook	Temperature	Humidity	Windy	Play golf
$\mathbf{x}_1$	Rainy	Hot	High	False	No
$\mathbf{x}_2$	Rainy	Hot	High	True	No
$\mathbf{x}_3$	Overcast	Hot	High	False	Yes
$\mathbf{x}_4$	Sunny	Mild	High	False	Yes
$\mathbf{x}_5$	Sunny	Cool	Normal	False	Yes
$\mathbf{x}_6$	Sunny	Cool	Normal	True	No
$\mathbf{x}_7$	Overcast	Cool	Normal	True	Yes
$\mathbf{x}_8$	Rainy	Mild	High	False	No
$\mathbf{x}_9$	Rainy	Cool	Normal	False	Yes
$\mathbf{x}_{10}$	Sunny	Mild	Normal	False	Yes
$\mathbf{x}_{11}$	Rainy	Mild	Normal	True	Yes
$\mathbf{x}_{12}$	Overcast	Mild	High	Ture	Yes
$\mathbf{x}_{13}$	Overcast	Hot	Normal	False	Yes
$\mathbf{x}_{14}$	Sunny	Mild	High	True	No

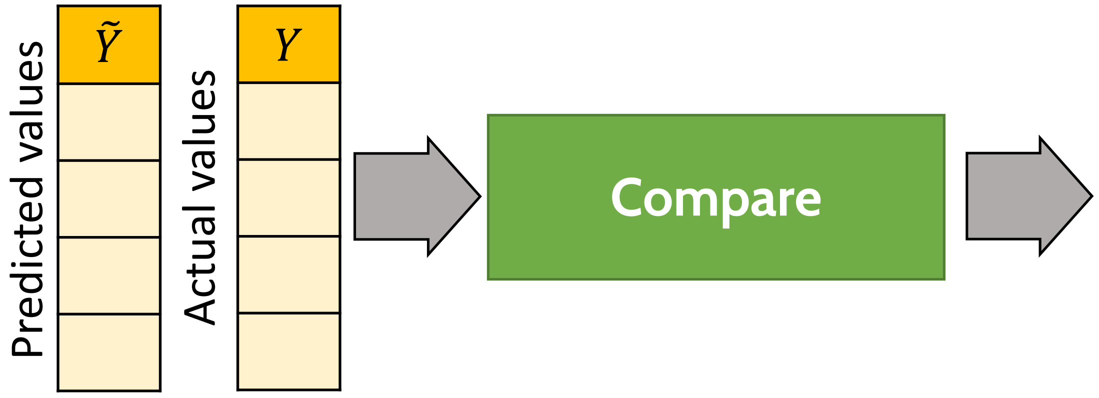
Classification Assessment

Classification Analysis



Classification Assessment

Classification Analysis



Actual values

Predicted values	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Confusion matrix

true positives (TP)
true negatives (TN)
false positives (FP)
false negatives (FN)

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{\text{Total}}$$

$$\text{Misclassification Rate} = \frac{(\text{FP} + \text{FN})}{\text{Total}} = 1 - \text{Accuracy}$$

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

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

Classification Assessment
Classification Analysis

Example

Actual values (columns) / Predicted values (rows)

	setosa	versicolor	virginica
setosa	**10**	2	4
versicolor	1	**16**	1
virginica	0	2	**9**

$$\text{Accuracy} = \frac{(10+16+9)}{45} = \frac{35}{45} = 0.78$$

$$\text{Misclassification Rate} = 1 - 0.78 = 0.22$$

$$\text{Recall}_{\text{setosa}} = \frac{10}{10+1+0} = \frac{10}{11} = 0.91$$

$$\text{Precision}_{\text{setosa}} = \frac{10}{10+2+4} = \frac{10}{16} = 0.625$$

$$\text{Recall}_{\text{versicolor}} = \frac{16}{2+16+2} = \frac{16}{20} = 0.8$$

$$\text{Precision}_{\text{versicolor}} = \frac{16}{1+16+1} = \frac{16}{18} = 0.89$$

$$\text{Recall}_{\text{virginica}} = ?$$

$$\text{Precision}_{\text{virginica}} = ?$$

Classification Assessment

Classification Analysis

Example

	Actual values: Cat	Actual values: Dog
Predicted values: Cat	**5**	2
Predicted values: Dog	3	**3**

$$\text{Accuracy} = \frac{(5+3)}{13} = \frac{8}{13} = 0.62$$

$$\text{Misclassification Rate} = \frac{(2+3)}{13} = \frac{5}{13} = 0.38$$

$$\text{Recall} = \frac{5}{5+3} = \frac{5}{8} = 0.625$$

$$\text{Precision} = \frac{5}{5+2} = \frac{5}{7} = 0.714$$

Regression analysis

Independent variable

Dependent variable

D	X_1	X_2	…	X_{10}	Y
$\mathbf{x}_1$					
$\mathbf{x}_2$					
$\mathbf{x}_3$					
…					
$\mathbf{x}_l$					
$\mathbf{x}_{l+1}$					
$\mathbf{x}_{l+2}$					
…					
$\mathbf{x}_n$					

For regression analysis

- The value we want to predict is **numeric data.**
- Known as **Dependent variable**

Example

- We know quantities of water and fertilizer providing to a tree for a month
- We want to predict the growth rate (height) of the tree.

Height?

Further Study

- **Book**:
  - Zaki, M., & Meira, W. (2014). Data mining and analysis : Fundamental concepts and algorithms. New York: Cambridge University Press.
  - Enders, C. (2010), Applied Missing Data Analysis. New York: Guilford Press.
- **Online lesson**: วิทยาการข้อมูลเบื้องต้น (Introduction to Data Science) – CMU MOOC https://thaimooc.org/courses/course-v1:CMU-MOOC+cmu034+2019_T1/about