

Crafted by :

**Raudhoh Fitra H.**

Training series number

**XXX.XXX.XX.XXX.XX**

Updated

**Q2.2020**

# Machine Learning I —

# — Classification Algorithms

# Classification Algorithm

## **Learning of binary classification**

- Given: a set of  $m$  examples  $(x_i, y_i)$   $i = 1, 2, \dots, m$  sampled from some distribution  $D$ , where  $x_i \in \mathbb{R}^n$  and  $y_i \in \{-1, +1\}$
- Find: a function  $f: \mathbb{R}^n \rightarrow \{-1, +1\}$  which classifies 'well' examples  $x_j$  sampled from  $D$ .

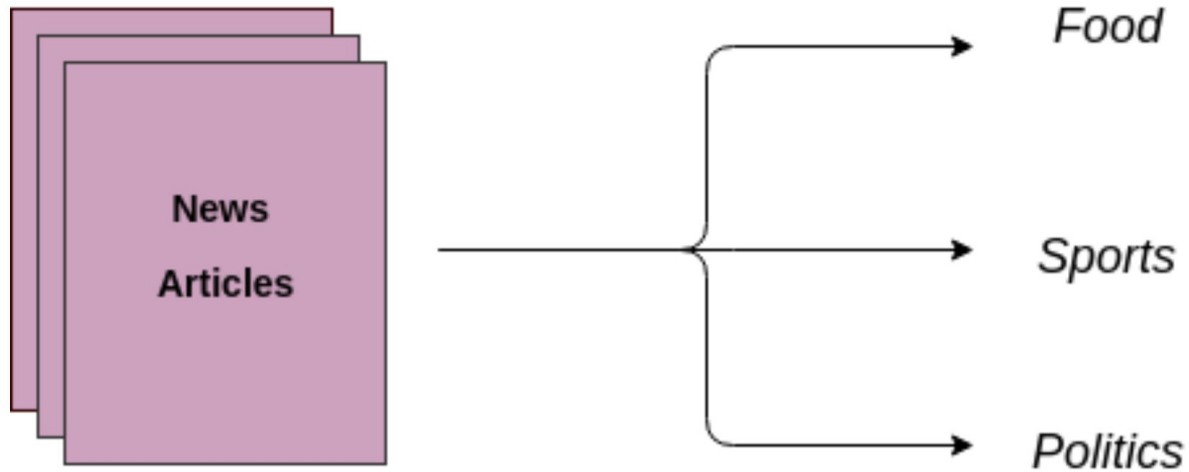
## **comments**

- The function  $f$  is usually a statistical model, whose parameters are learnt from the set of examples.
- The set of examples are called – 'training set'.
- $Y$  is called – 'target variable', or 'target'.
- Examples with  $y_i = +1$  are called 'positive examples'. Examples with  $y_i = -1$  are called 'negative examples'.

## Real World Implementation

- Text categorization: spam detection
- Face detection: Signature recognition: Customer discovery
- Medicine: Predict if a patient has heart ischemia by a spectral analysis of his/her ECG.

## Text Categorization



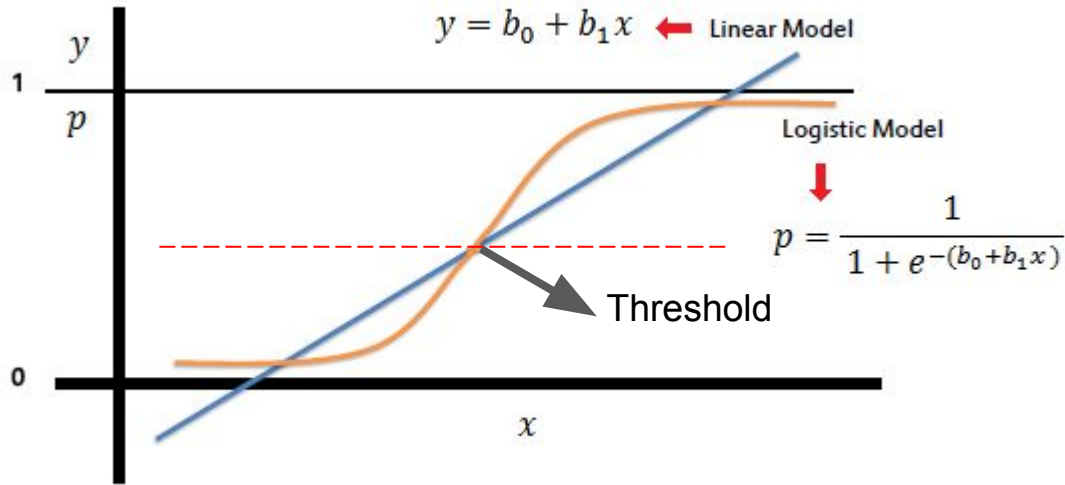
## Face Detection



# Logistic Regression

# Classification Method

, output :



} Class 0  
} Class 1



# Logistic Regression Assumption

1. Binary logistic regression requires the dependent variable to be binary
2. For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome
3. Only the meaningful variables should be included
4. The independent variables should be independent of each other. That is, the model should have little or no multicollinearity
5. The independent variables are linearly related to the log odds.
6. Logistic regression requires quite large sample sizes.

# Naive Bayes



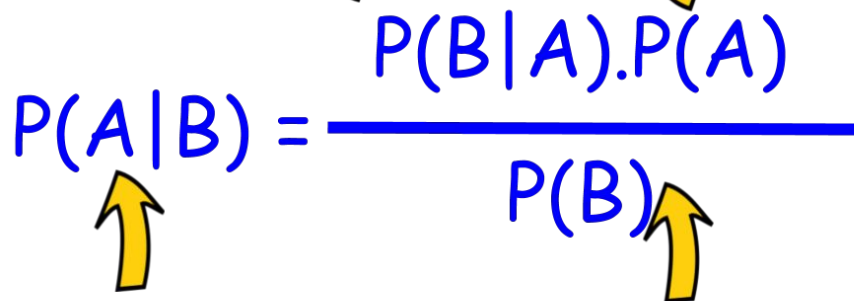
# Bayes Theorem

## LIKELIHOOD

The probability of "B" being True, given "A" is True

## PRIOR

The probability "A" being True. This is the knowledge.



The diagram shows the Bayes Theorem formula with four yellow arrows pointing to its components: one from 'LIKELIHOOD' to  $P(B|A)$ , one from 'PRIOR' to  $P(A)$ , one from 'POSTERIOR' to  $P(A|B)$ , and one from 'MARGINALIZATION' to  $P(B)$ .

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

## POSTERIOR

The probability of "A" being True, given "B" is True

## MARGINALIZATION

The probability "B" being True.

$$\textit{Posterior} = \frac{\textit{Likelihood} \times \textit{Prior}}{\textit{Evidence}}$$

**prior** = probabilitas suatu kejadian  $p(h_j)$ ,  $j = 1, 2$

**likelihood** = probabilitas kejadian bersyarat

Misal,  $p(x|h_j)$ , peluang atribut bernilai x jika diketahui h

**evidence** =  $\sum_{j=1}^2 p(x|h_j)p(h_j)$

# Case Study

Terdapat data diagnosis penyakit dengan 2 alternatif hipotesis :

- (1) Pasien mengidap kanker
- (2) Pasien tidak mengidap kanker

Uji lab menghasilkan keluaran positif dimana memang benar 98% dari semua kasus mengidap kanker, dan keluaran negatif sebesar 97% dimana memang pasien tidak mengidap kanker.

Dari situasi tersebut bisa disimpulkan :

$$\begin{aligned} P(\text{kanker}) &= 0,008 & ; & P(\sim\text{kanker}) = 0,992 \\ P(+ | \text{kanker}) &= 0,98 & ; & P(- | \text{kanker}) = (1-0,98) = 0,02 \\ P(+ | \sim\text{kanker}) &= 0,03 & ; & P(- | \sim\text{kanker}) = 0,97 \end{aligned}$$

Jika terdapat uji lab baru yang mana hasilnya positif, apakah hipotesis yang paling mungkin untuk pasien tersebut?

$$P(\text{kanker} | +) = P(+ | \text{kanker}) P(\text{kanker}) = (0,98)(0,008) = 0,0078$$

$$P(\sim\text{kanker} | +) = P(+ | \sim\text{kanker}) P(\sim\text{kanker}) = (0,03)(0,992) = 0,0298$$

Probabilitas posterior  $P(\text{kanker} | +) = 0,0078 / (0,0078 + 0,0298) = 0,21$

Maka pasien tersebut diprediksi **tidak menderita kanker**.

$$p(kanker \mid atr_1, atr_2, \dots, atr_n) = \frac{p(atr_1, atr_2, \dots, atr_n) \times p(kanker)}{p(atr_1, atr_2, \dots, atr_n)}$$

Atr<sub>1</sub> misal merokok, atr<sub>2</sub> misal umur, atau pekerjaan.

# Naive Bayes

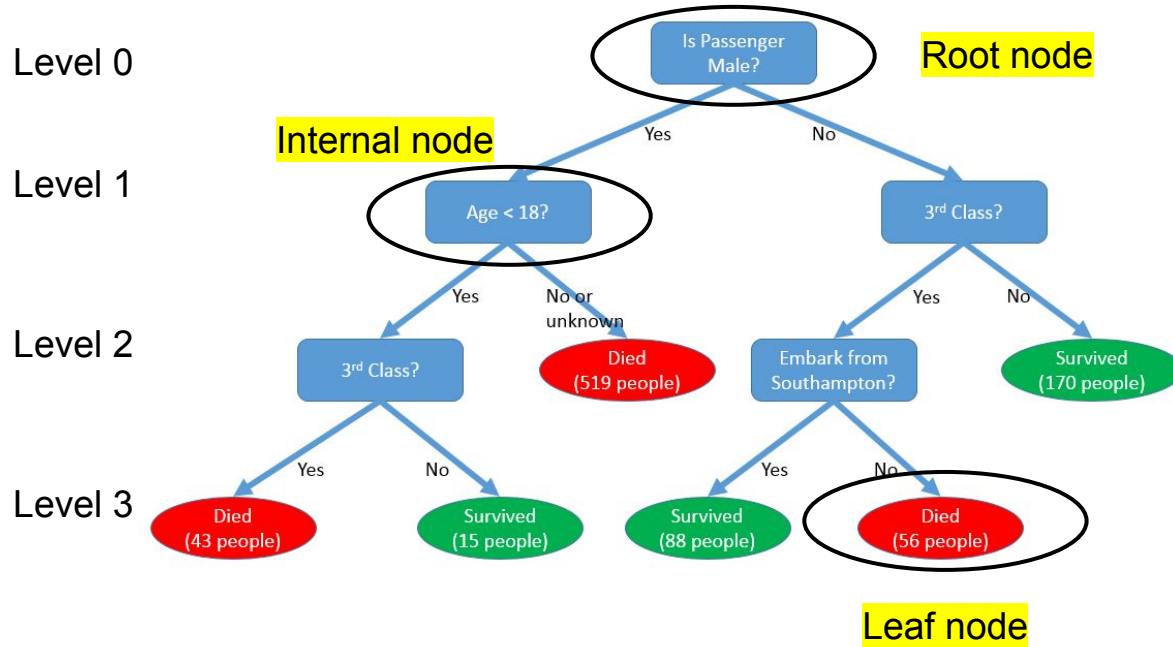
Pelanggan	Kartu	Panggilan	Blok	Bonus
Andi	Prabayar	Sedikit	Sedang	Tidak
Budi	Pascabayar	Banyak	Sedang	Ya
Citra	Prabayar	Banyak	Sedang	Ya
Dedi	Prabayar	Banyak	Rendah	Tidak
Evan	Pascabayar	Cukup	Tinggi	Ya
Feni	Pascabayar	Cukup	Sedang	Ya
Gito	Prabayar	Cukup	Sedang	Ya
Hani	Prabayar	Cukup	Rendah	Tidak
Jodi	Pascabayar	Sedikit	Tinggi	Ya
Kafi	Pascabayar	Banyak	Tinggi	Ya
Linda	Pascabayar	Sedikit	Rendah	Ya

Fernando	Pascabayar	Cukup	Rendah	?
----------	------------	-------	--------	---





# | Decision Tree



- Bivariate, or multivariate classification
- Have ability to sort the features based on feature importance
- Discrete output
- Transparent outcome
- Robust

# Information Gain

# Sample case

Outlook	Temperature	Humidity	Wind	Played football(yes/no)
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

**Playing football or not?**

# Confusion Matrix

## Type I Error



## Type II Error





# Confusion Matrix

## Two classes

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

## More than 2 classes

		Predicted Values				total
		Null'	Low'	Moderate'	High'	
Real Values	Null	True Null	False Low	False Moderate	False High	N
	Low	False Null	True Low	False Moderate	False High	L
	Moderate	False Null	False Low	True Moderate	False High	M
	High	False Null	False Low	False Moderate	True High	H
total		N'	L'	M'	H'	

		Predicted/Classified	
		Negative	Positive
Actual	Negative	998	0
	Positive	1	1



## Precision

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

## Recall

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

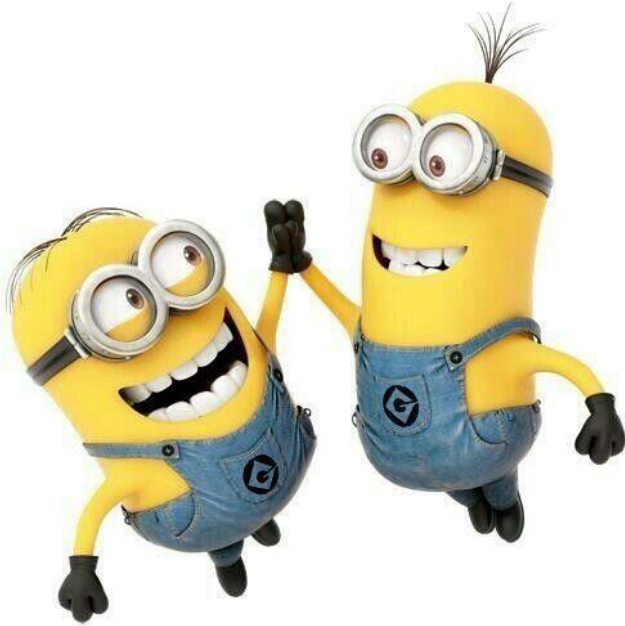
		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

## Specificity

		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$



Let's Practice, yeay!



# Thank You