Supervised Learning
Classification
&
Naïve Bayes

Slides นี้ได้นำเนื้อหาบางส่วนจากวิชา 0904 413 โดย **อาจารย์อนุพงศ์ สุขประเสริฐ,** CS583 Bing Liu, UIC และ Gil, Yolanda (Ed.) Introduction to Computational Thinking and Data Science. Available from http://www.datascience4all.org (และได้ปรับเปลี่ยนเนื้อหาบางส่วน)

Supervised learning (การเรียนรู้แบบมีผู้สอน)

- การเรียนรู้แบบมีผู้สอน คือการทำให้คอมพิวเตอร์สามารถเรียนรู้หาคำตอบของปัญหาได้ด้วยตัวเอง หลังจากเรียนรู้จากชุดข้อมูลตัวอย่างไปแล้ว

การจำแนกข้อมูล (Classification)

- การจำแนกข้อมูล (Classification) หมายถึง การจำแนกหรือแบ่งประเภทข้อมูล โดยหาต้นแบบหรือสำรวจจุดเด่นจุดด้อยที่ปรากฏอยู่ภายในชุดข้อมูล โดยใช้ข้อมูลที่มีอยู่จำนวนหนึ่งในการสร้างต้นแบบ (training data)

- ตัวแบบที่ได้รับนั้น จะสามารถนำไปใช้ในการกำหนดประเภทของชุดข้อมูลว่ามีกี่ประเภท อะไรบ้าง อย่างเหมาะสม เพื่อใช้ทำนายประเภทของข้อมูลใหม่ วัตถุ (เรียกว่า การแบ่งประเภท -- classification) ที่ไม่เคยเห็นมาก่อน (unseen data) ให้อยู่ตามประเภทหรือหมวดหมู่ที่เหมาะสม

จุดประสงค์ของการจำแนกประเภทข้อมูล

- คือการสร้างโมเดลการแยกแอทริบิวท์หนึ่งโดยขึ้นกับแอทริบิวท์อื่น โมเดลที่ได้จากการจำแนกประเภทข้อมูลจะทำให้สามารถพิจารณาคลาสในข้อมูลที่ยังมิได้แบ่งกลุ่มในอนาคตได้
- เทคนิคการจำแนกประเภทข้อมูลนี้ได้นำไปประยุกต์ใช้ในหลายด้าน เช่น การจัดกลุ่มลูกค้าทางการตลาด, การตรวจสอบความผิดปกติ และการวิเคราะห์ทางการแพทย์ เป็นต้น

An example: data (loan application)

Approved or not

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

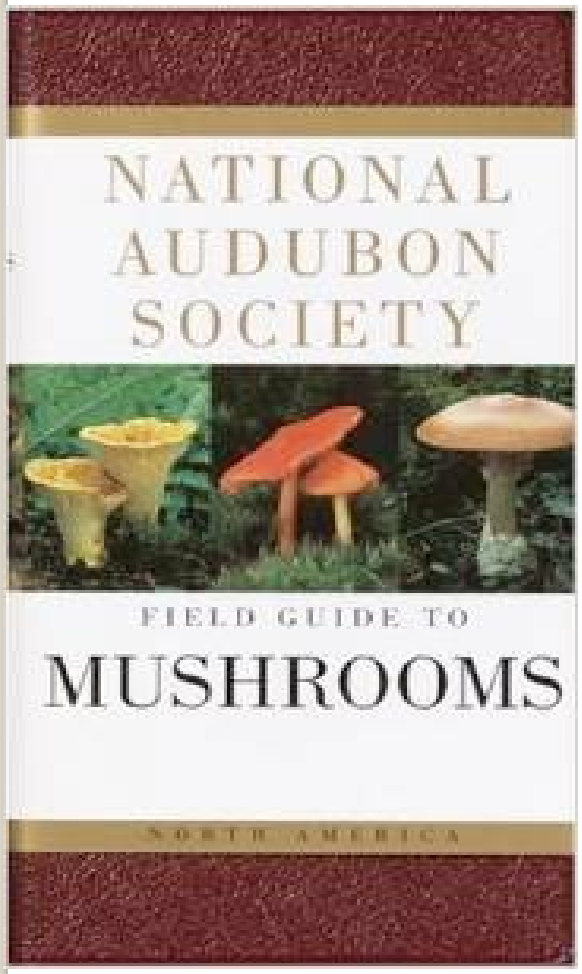
- ทำนายข้อมูลใหม่ว่าจะอนุมัติให้กู้เงินได้หรือไม่ Yes (approved) and No (not approved)
- Use the model to classify future loan applications into

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

4

Classifying Mushrooms

- เห็ดแบบไหนมีพิษหรือไม่มีพิษ



https://archive.ics.uci.edu/ml/datasets/Mushroom

Classifying Iris Plants

- จำแนกประเภทพันธุ์ดอกไม้

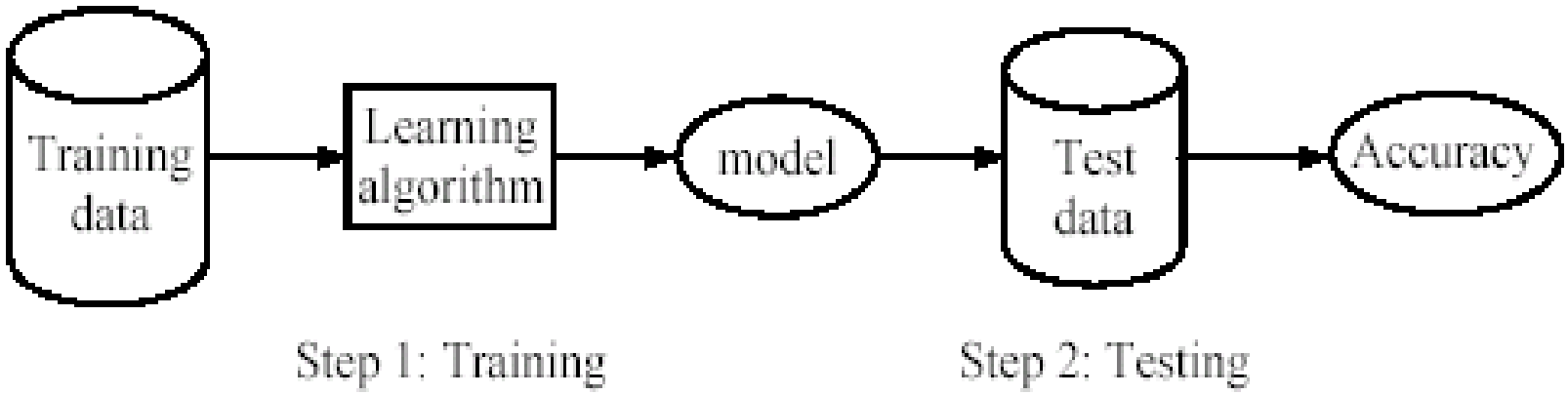
- Iris Setosa
- Iris Versicolour
- Iris Virginica

https://archive.ics.uci.edu/ml/datasets/Iris

Supervised learning process: two steps

- Learning (training): Learn a model using the training data
- Testing: Test the model using unseen test data to assess the model accuracy

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a (tp)	b (fn)
	Class=No	c (fp)	d (tn)

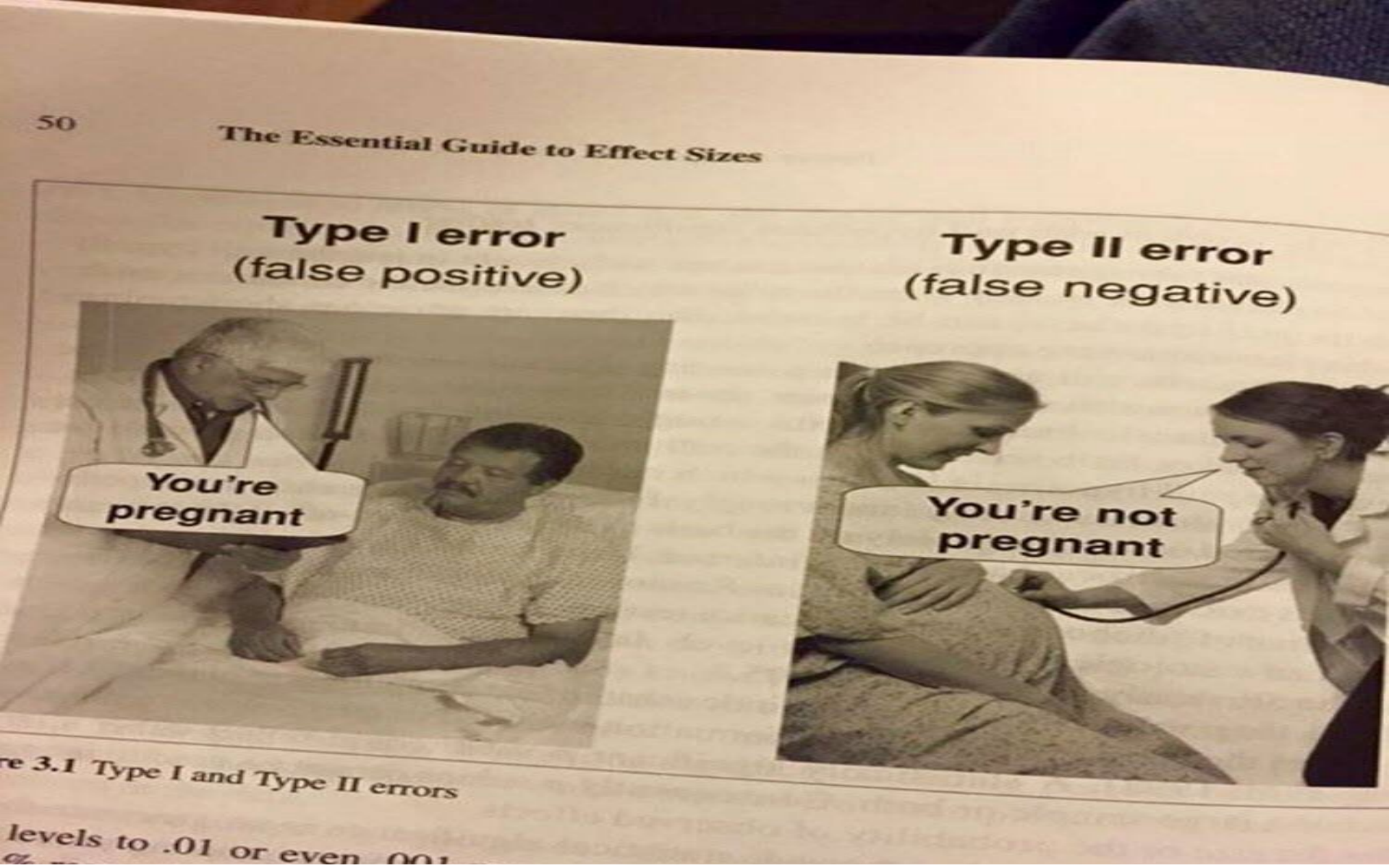
a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$



[http://imgur.com/5vTarFz]

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Classifier Evaluation Metrics: Confusion Matrix

Confusion Matrix:

Actual class\Predicted class	C_1	¬ C_1
C_1	**True Positives (TP)**	**False Negatives (FN)**
¬ C_1	**False Positives (FP)**	**True Negatives (TN)**

Example of Confusion Matrix:

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

เทคนิคการการจำแนกข้อมูล (Classification Techniques)

- เทคนิคการจำแนกประเภทข้อมูลเป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้จากกลุ่มตัวอย่างข้อมูลที่เรียกว่าข้อมูลสอนระบบ (training data)

- แต่ละแถวของข้อมูลประกอบด้วยฟิลด์หรือแอทริบิวท์จำนวนมาก แอทริบิวท์นี้อาจเป็นค่าต่อเนื่อง (continuous) หรือค่ากลุ่ม (categorical) โดยจะมีแอทริบิวท์แบ่ง (classifying attribute) ซึ่งเป็นตัวบ่งชี้คลาสของข้อมูล

- เทคนิคในการจำแนกกลุ่มข้อมูลด้วยคุณลักษณะต่างๆที่ได้มีการกำหนดไว้แล้วสร้างแบบจำลองเพื่อการพยากรณ์ค่าข้อมูล (Predictive Model) ในอนาคต เรียกว่า Supervised learning ซึ่งได้แก่
  - Decision Tree
  - Naive Bayes
  - K-Nearest Neighbors (kNN)
  - Linear Regression
  - Neural Network

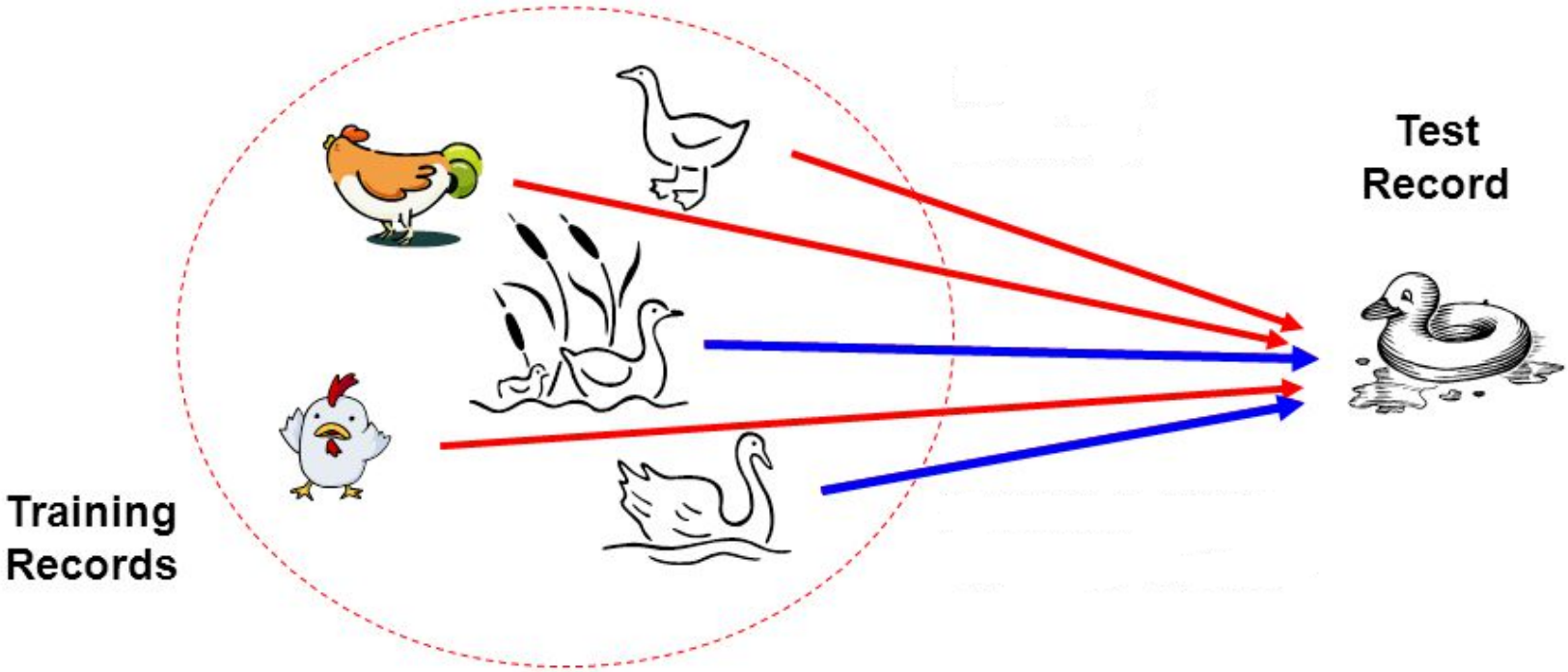
Naive Bayes

- ใช้หลักการความน่าจะเป็น **(Probability)** ทางทฤษฎีสถิติ
- โอกาสที่เกิดเหตุการณ์จากเหตุการณ์ทั้งหมด ใช้สัญลักษณ์ P() หรือ Pr() เช่น
  - การโยนเหรียญความน่าจะเป็นของการเกิดหัวและก้อย
    - โอกาสที่จะออกหัว มีความน่าจะเป็น ½ = 0.5
    - โอกาสที่จะออกก้อย มีความน่าจะเป็น ½ = 0.5
  - ความน่าจะเป็นของการพบ spam email
    - มี email ทั้งหมด 100 ฉบับ
    - มี spam email ทั้งหมด 20 ฉบับ
    - มี normal email ทั้งหมด 80 ฉบับ
    - โอกาสที่ email จะเป็น spam มีความน่าจะเป็น 20/100 = 0.2 หรือ P(spam) = 0.2
    - โอกาสที่ email จะเป็น normal มีความน่าจะเป็น 80/100 = 0.8 หรือ P(normal) = 0.8

Thomas Bayes
1702 - 1761

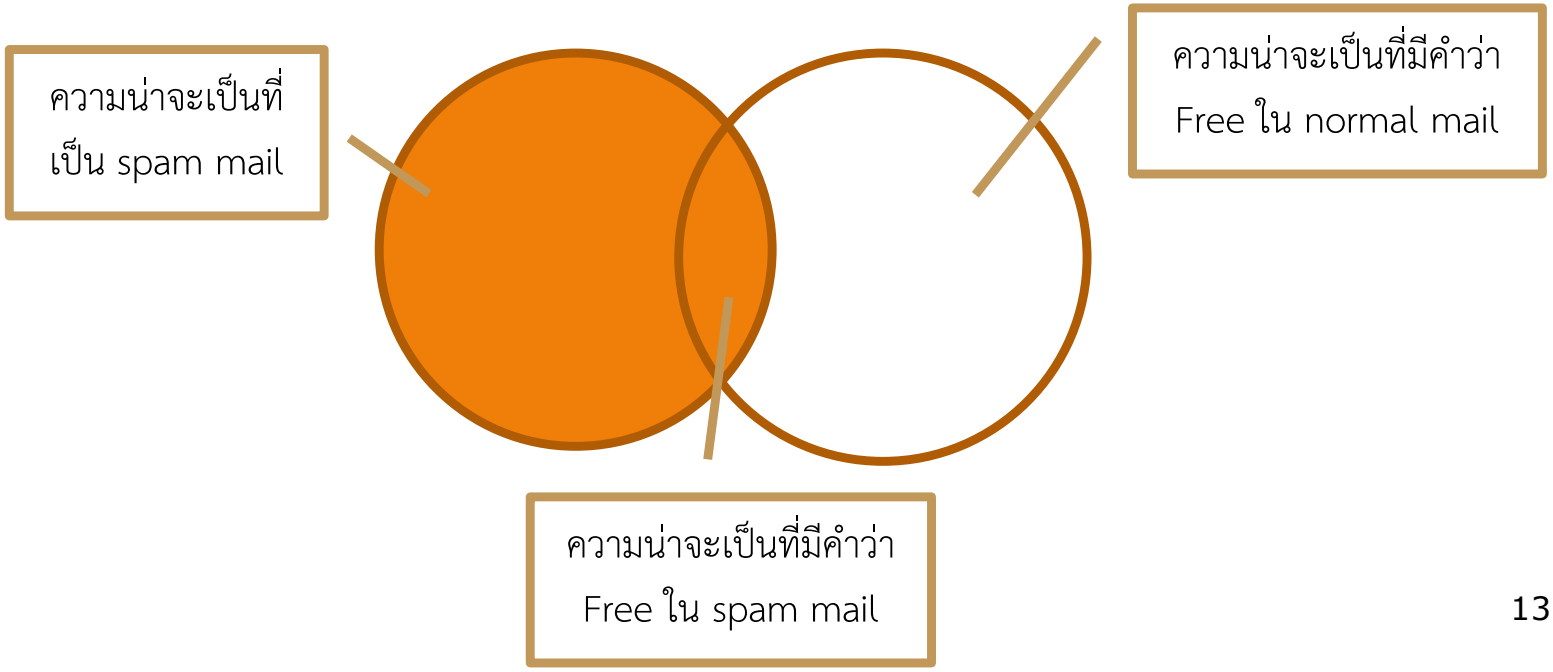
Naïve Bayes Classifier

- Principle
  - If it walks like a duck, quacks like a duck, then it is probably a duck



Probability

- **Joint Probability** คือ ความน่าจะเป็นของ 2 เหตุการณ์ที่เกิดขึ้นร่วมกัน
- ตัวอย่าง: ความน่าจะเป็นที่มีคำว่า Free อยู่ใน spam mail
- สัญลักษณ์ P(Free=Y ∩ spam)



Naive Bayes

- ใช้หลักการความน่าจะเป็น (probability)

ความน่าจะเป็นที่ B เกิดก่อนและ A เกิดตามมา

ความน่าจะเป็นที่ A และ B เกิดร่วมกัน

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

- P(A|B) คือ ค่า conditional probability หรือค่าความน่าจะเป็นที่เกิดเหตุการณ์ B ขึ้นก่อนและจะมีเหตุการณ์ A ตามมา
- P(A ∩ B) คือ ค่า joint probability หรือค่าความน่าจะเป็นที่เหตุการณ์ A และเหตุการณ์ B เกิดขึ้นร่วมกัน
- P(B) คือ ค่าความน่าจะเป็นที่เหตุการณ์ B เกิดขึ้น

- ในลักษณะเดียวกันเราจะเขียน $P(B|A)$ หรือค่าความน่าจะเป็นที่เหตุการณ์ A เกิดขึ้นก่อนและเหตุการณ์ B เกิดขึ้นตามมาทีหลังได้เป็น

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

Naive Bayes

- จากทั้ง 2 แบบจะเห็นว่ามีค่า P(A ⋂ B) ที่เหมือนกันอยู่ดังนั้นสามารถเขียนสมการของ P(A ⋂ B) ได้เป็นดังนี้

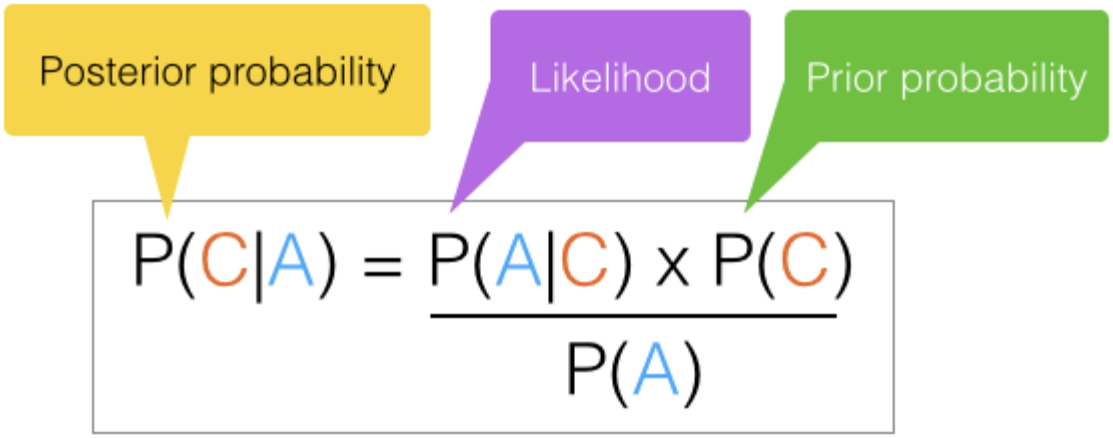
$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$$

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

Bayes Theorem

- สมการข้างต้นเรียกว่า Bayes theorem หรือทฤษฎีของเบย์ ในการนำไปใช้งานทางด้านการจำแนกข้อมูล (classification) ซึ่งเมื่อเอามาประยุกต์ใช้สามารถ represent สัญลักษณ์ A และ B โดยที่ A คือ แอตทริบิวต์ (attribute) และ C คือ ค่าคลาส (class)

Naive Bayes



จากสมการของ Bayes จะมี 3 ส่วนที่สำคัญ คือ

- **Posterior probability** หรือ P(C|A) คือ ค่าความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น A จะมีคลาส C
- **Likelihood** หรือ P(A|C) คือ ค่าความน่าจะเป็นที่ข้อมูล training data ที่มีคลาส C และมีแอตทริบิวต์ A โดยที่ $A = a_1 \cap a_2 \ldots \cap a_M$ โดยที่ M คือจำนวนแอตทริบิวต์ใน training data
- **Prior probability** หรือ P(C) คือ ค่าความน่าจะเป็นของคลาส C

Naive Bayes

- แต่การที่แอตทริบิวต์ $A = a_1 \cap a_2 \ldots \cap a_M$ ที่เกิดขึ้นใน training data อาจจะมีจำนวนน้อยมากหรือไม่มีรูปแบบของแอตทริบิวต์แบบนี้เกิดขึ้นเลย ดังนั้นจึงได้ใช้หลักการที่ว่าแต่ละแอตทริบิวต์เป็น independent ต่อกันทำให้สามารถเปลี่ยนสมการ $P(A|C)$ ได้เป็น

$$P(A|C) = P(a_1|C) \times P(a_2|C) \times \ldots P(a_M|C)$$

- วิธีการคำนวณค่าต่างๆ จากไฟล์ training data เพื่อสร้างเป็นโมเดล Naive Bayes

ตัวอย่าง 1 การคำนวณ training data เพื่อสร้างเป็นโมเดล Naive Bayes

Class:
C1:buys_computer = 'yes'
C2:buys_computer = 'no'

New Data
X = (age <=30,
Income = medium,
Student = yes
Credit_rating = fair)

Buy_computer ???

age	income	student	credit_rating	com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

ตัวอย่าง 1 การคำนวณ training data เพื่อสร้างเป็นโมเดล Naive Bayes

- $P(C_i)$: P(buys_computer = "yes") = 9/14 = 0.643
  P(buys_computer = "no") = 5/14= 0.357

- Compute $P(X|C_i)$ for each class
  P(age = "<=30" | buys_computer = "yes") = 2/9 = 0.222
  P(age = "<= 30" | buys_computer = "no") = 3/5 = 0.6
  P(income = "medium" | buys_computer = "yes") = 4/9 = 0.444
  P(income = "medium" | buys_computer = "no") = 2/5 = 0.4
  P(student = "yes" | buys_computer = "yes) = 6/9 = 0.667
  P(student = "yes" | buys_computer = "no") = 1/5 = 0.2
  P(credit_rating = "fair" | buys_computer = "yes") = 6/9 = 0.667
  P(credit_rating = "fair" | buys_computer = "no") = 2/5 = 0.4

- **X = (age <= 30 , income = medium, student = yes, credit_rating = fair)**

$P(X|C_i)$: P(X|buys_computer = "yes") = 0.222 × 0.444 × 0.667 × 0.667 = 0.044
P(X|buys_computer = "no") = 0.6 × 0.4 × 0.2 × 0.4 = 0.019

$P(X|C_i)*P(C_i)$: P(X|buys_computer = "yes") * P(buys_computer = "yes") = 0.028
P(X|buys_computer = "no") * P(buys_computer = "no") = 0.007

Therefore, X belongs to class ("buys_computer = yes")

ตัวอย่าง 2 การคำนวณ training data เพื่อสร้างเป็นโมเดล Naive Bayes

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	**No**
2	No	Married	100K	**No**
3	No	Single	70K	**No**
4	Yes	Married	120K	**No**
5	No	Divorced	95K	**Yes**
6	No	Married	60K	**No**
7	Yes	Divorced	220K	**No**
8	No	Single	85K	**Yes**
9	No	Married	75K	**No**
10	No	Single	90K	**Yes**

- Class: $P(C) = N_c/N$
  - e.g., P(No) = 7/10,
    P(Yes) = 3/10
- For discrete attributes:

$$P(A_i \mid C_k) = |A_{ik}|/ N_c$$

  - where $|A_{ik}|$ is number of instances having attribute A_i and belongs to class C_k
  - Examples:

P(Status=Married|No) = 4/7
P(Refund=Yes|Yes)=0

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

ตัวอย่าง 2 การคำนวณ training data เพื่อสร้างเป็นโมเดล Naive Bayes

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- Normal distribution: $P(x: \mu, \sigma^2)$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - One for each (A_i, c_i) pair

- For (Income, Class=No):
  - If Class=No
    - sample mean = 110
    - sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

ตัวอย่าง 2 การคำนวณ training data เพื่อสร้างเป็นโมเดล Naive Bayes

Given a Test Record or New Data:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

- Normal distribution:

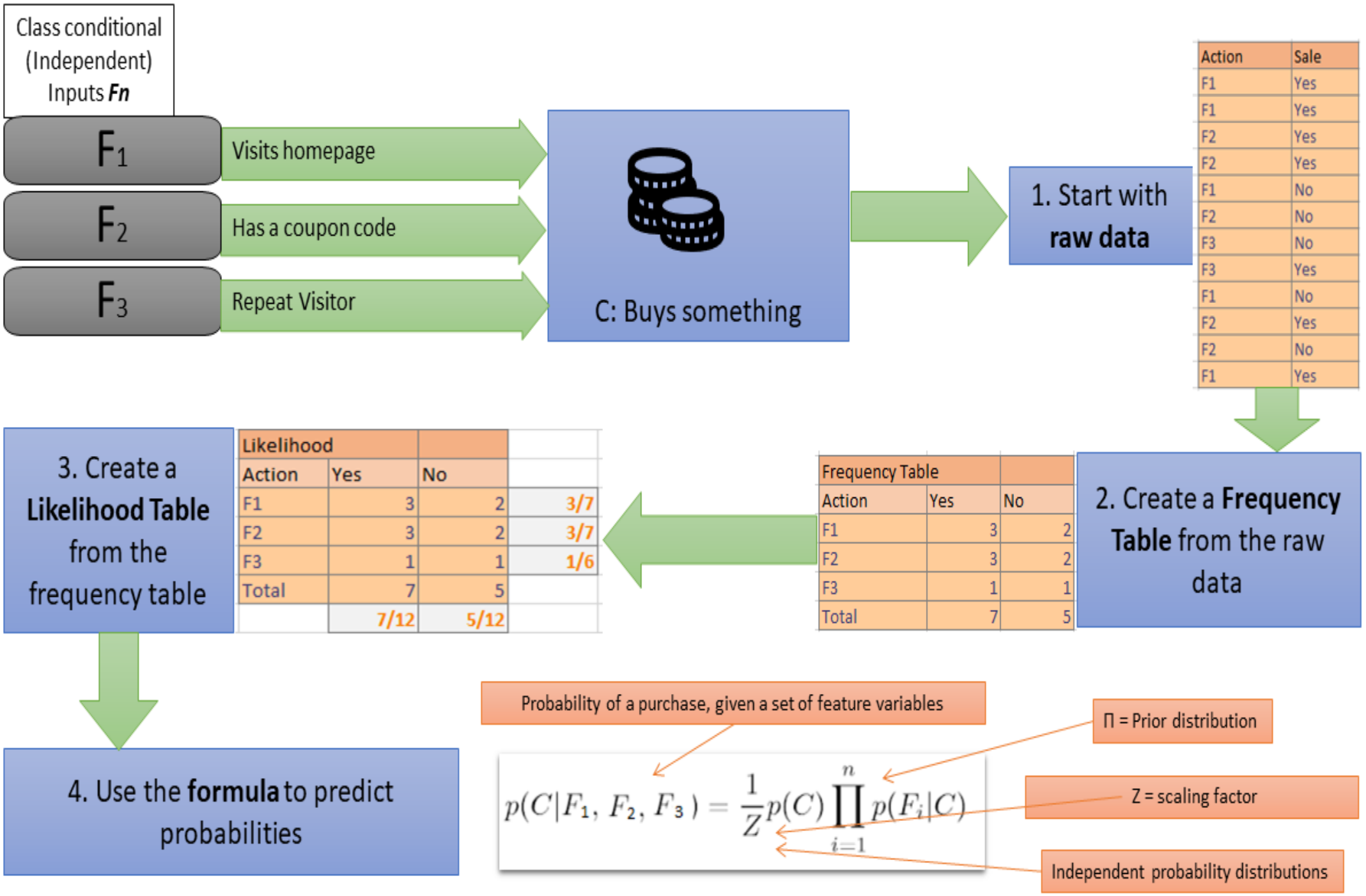
$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \quad P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Conclusion on Naïve Bayes classifiers

- Naïve Bayes is based on the independence assumption
  - Training is very easy and fast; just requiring considering each attribute in each class separately
  - Test is straightforward; just looking up tables or calculating conditional probabilities with normal distributions

- Naïve Bayes is a popular generative classifier model
  1. **Performance of naïve Bayes** is competitive to most of state-of-the-art classifiers even if in presence of violating the independence assumption
  2. It has many successful applications, e.g., spam mail filtering
  3. A good candidate of a base learner in ensemble learning
  4. Apart from classification, naïve Bayes can do more…

Q. If a person visits the homepage, has a coupon code, and is a repeat visitor, what is the probability they will buy something?



https://www.datasciencecentral.com/profiles/blogs/naive-bayes-in-one-picture-1