樸素貝式分類法
處理數值型條件屬性

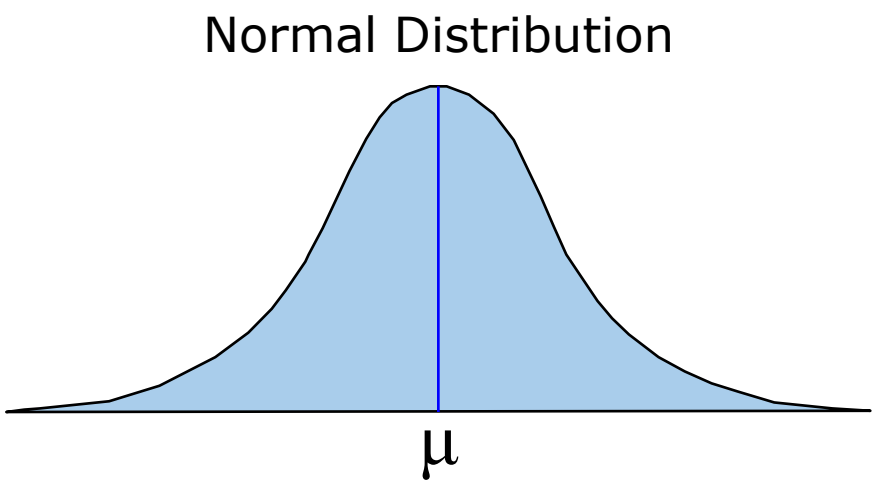
國立宜蘭大學資訊工程系
吳政瑋 助理教授
silvemoonfox@hotmail.com

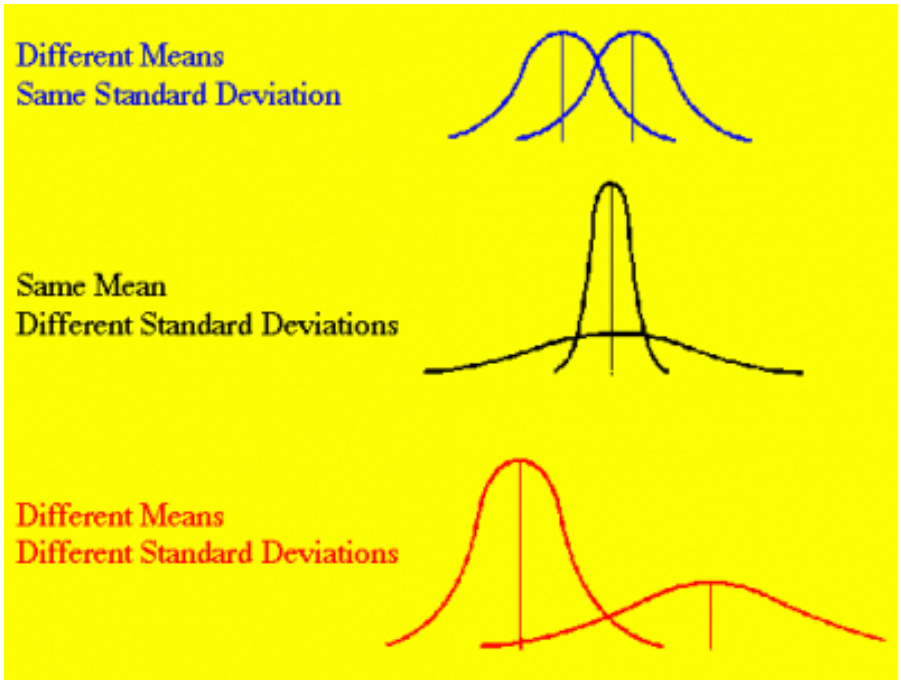
Example Dataset (逾期還款資料集)

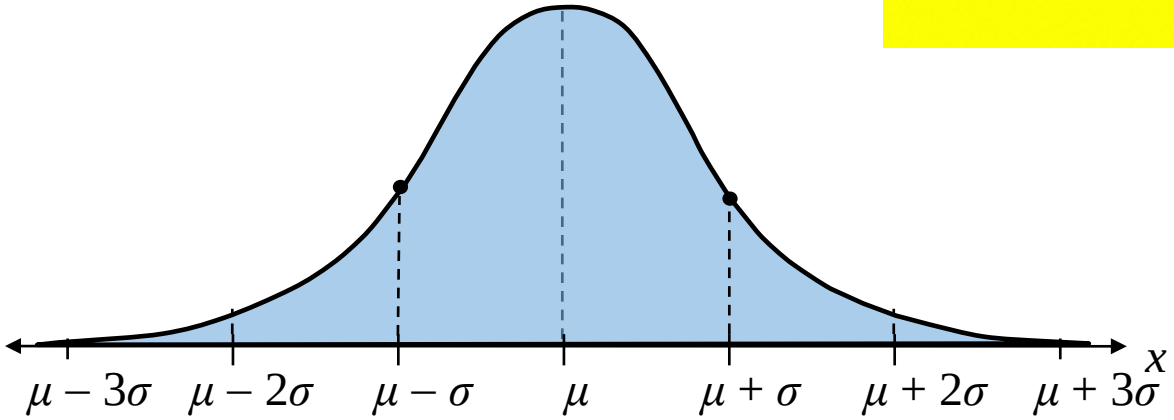
- 此資料集紀錄借貸者的相關資料，其條件屬性中，**Annual Income(年收入)** 數值型屬性。

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

常態分佈(Normal Distribution)
平均值與標準差之關係







機率密度函數 (Probability Density Function；PDF)

- 給一個**平均值 μ** 以及**標準差 σ**，我們可畫出一個常態分佈。
- 可以根據下列的**機率密度函式**公式，計算任意實數 x_i 發生在這個分佈的機率。

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

- https://www.geogebra.org/m/T5r6rJnj
- https://www.easycalculation.com/zh/statistics/standard-deviation.php

計算平均值、標準差及機率密度函數之線上工具

- 計算**平均值**及**標準差**之線上工具
  - https://www.geogebra.org/m/T5r6rJnj

- 計算**機率密度函數**之線上工具
  - https://www.geogebra.org/m/T5r6rJnj
  - https://www.danielsoper.com/statcalc/calculator.aspx?id=54

樸素貝式分類法
處理數值型條件屬性

- 對於會逾期還款的那群人(**目標屬性為 No** 的資料)，他們的Annual Income(年收入)之**平均數**及**標準差**為：

$$\bar{X}_{No} = \frac{125+100+...+75}{7} = 110$$

$$S_{No}^2 = \frac{(125-110)^2+(100-110)^2+...+(75-110)^2}{7} = 2975$$

$$S_{No} = 54.54$$

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

樸素貝式分類法
處理數值型條件屬性

- 假設現在有一筆資料，年收入為120K。
- 則在目標屬性值為No時，年收入120K屬於No這群人的條件機率即為：

$$P(X = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

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

樸素貝式分類法
處理數值型條件屬性

- 所有屬性在Yes及No的條件機率列表如下：

Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
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

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

For Annual Income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

樸素貝式分類法
處理數值型條件屬性

- 現在有一位顧客 X = {Home Owner = No, Marital Status = Married, Annual Income = 120K}則他是否會逾期還款(Defaulted Borrower =?)?

- 解題步驟
  - 先算 $P(B=No|X)$
  - 再算 $P(B=Yes|X)$
  - 比較 $P(B=No|X)$及 $P(B=Yes|X)$，機率大者勝

樸素貝式分類法
處理數值型條件屬性

- 現在有一位顧客 X = {Home Owner = No, Marital Status = Married, Annual Income = 120K}則他是否會逾期還款(Defaulted Borrower =?)?

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

$$P(B = No \mid H) = P(B = No)\prod_{i=1}^{3} P(H_i \mid B = No)$$

$$= P(No) \times P(Home\ Owner = No \mid No) \times P(Status = Married \mid No) \times P(Annual\ Income = 120K \mid No)$$

$$= 0.7 \times \frac{4}{7} \times \frac{4}{7} \times 0.0072$$

$$= 0.00165.$$

$$P(B = Yes \mid H) = P(B = Yes)\prod_{i=1}^{3} P(H_i \mid B = Yes)$$

$$= P(Yes) \times P(Home\ Owner = No \mid Yes) \times P(Status = Married \mid Yes) \times P(Annual\ Income = 120K \mid Yes)$$

$$= 0.3 \times 1 \times 0 \times 1.2 \times 10^{-9}$$

$$= 0.$$