

機器學習 02

機率分佈

楊智淵 2025/9/12

多項式曲線擬合

- 課本上開場的演算法
- 資料維度：1維，符號 t
- 資料筆數：10筆
- 它們是 $\sin(2\pi x)$ 加上雜訊合成出來的數據
- 現在我們的演算法假設他們應該用一條一元多次多項式來模擬

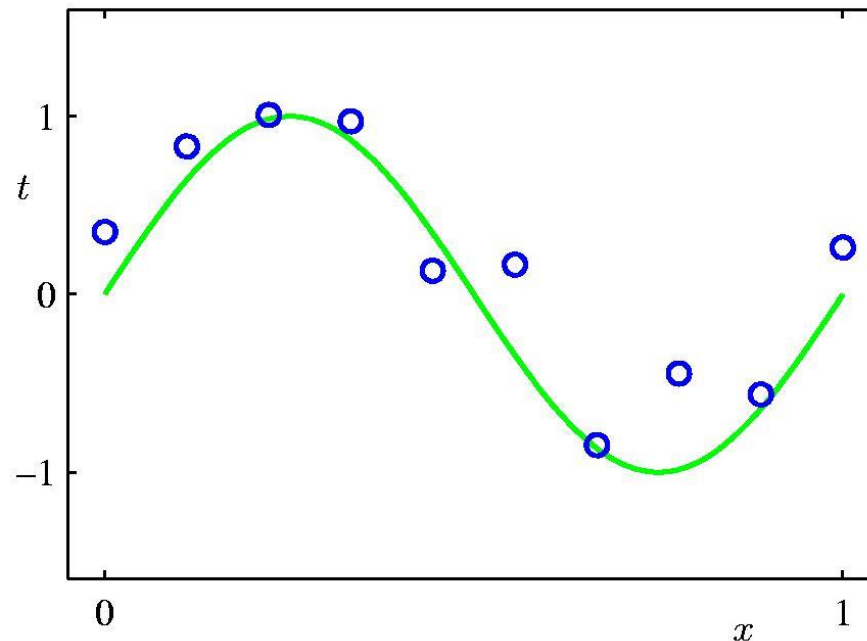


Figure 1.2

多項式的表示式

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j \quad (1.1)$$

- 回想一下高中學過的一元二次和三次多項式
- $ax^2 + bx + c$
- $ax^3 + bx^2 + cx + d$
- (1.1)式是一般表示式，有M次，共M+1個參數，這M+1個參數就是我們要機器學習的知識。這M+1個參數的值，會由你的資料決定。而M該是多少，你必需事先決定，所以M稱為超參數，而w是參數，用w這個符號因為我們要決定他的權重(weight)。

誤差平方和函式

- 你有了一個假設的模型之後，如何決定權重？
- 一個很直覺的想法，就是讓錯誤最小。
- 所以我們要為錯誤下數學式的定義。
- 平方和是很常用的錯誤度量法。它平等對待每一筆資料，以偏離的平方計算錯誤，寧可筆筆小偏誤，也不要看到一筆大偏誤。

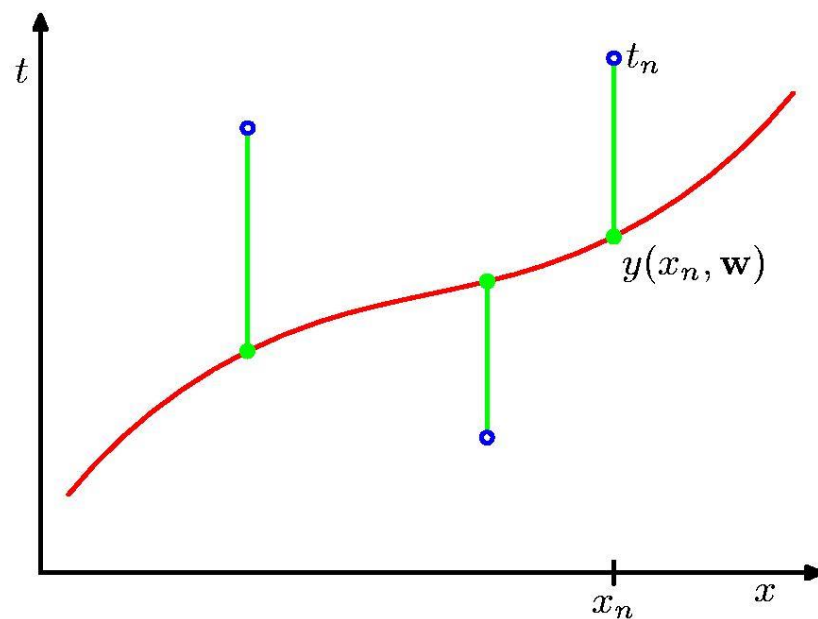


Figure 1.3

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 \quad (1.2)$$

當 $M=0$ 時

- 我們只有一個參數，演算法產生的值是定值，在這裏例子，是0.2左右，會讓10個資料點的錯誤最小。

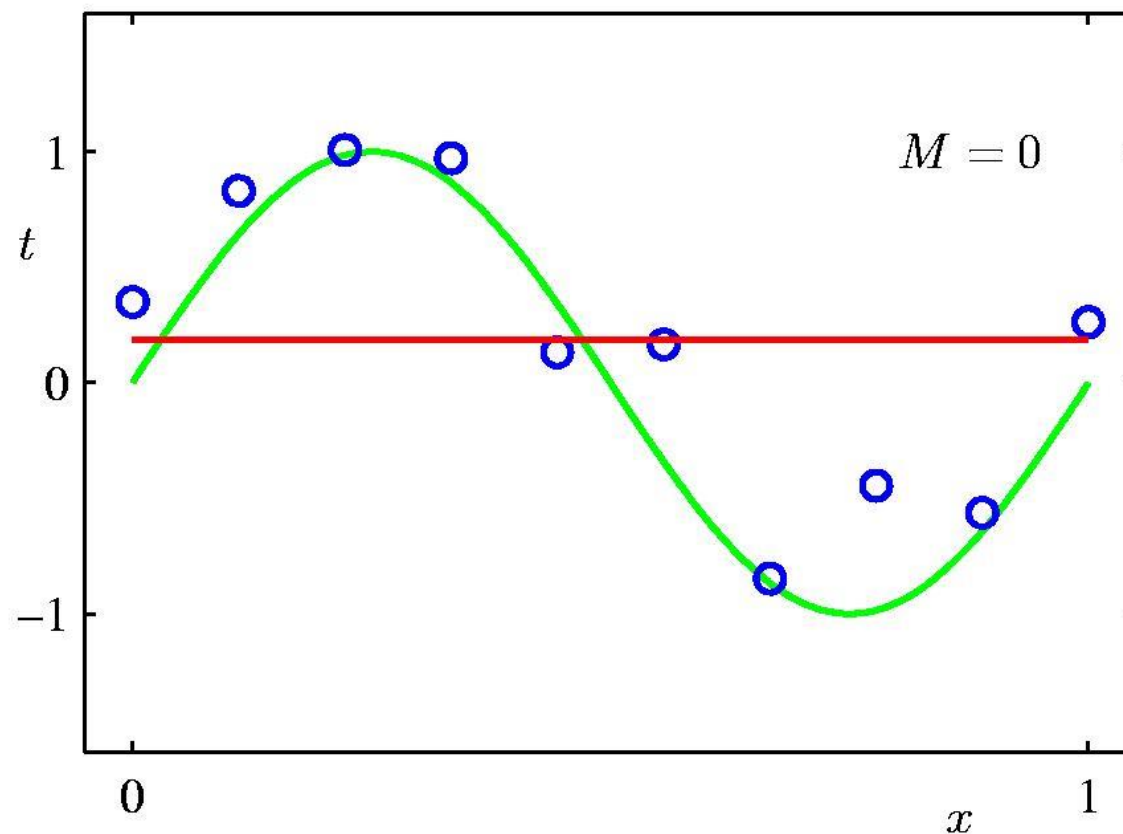


Figure 1.4

當 $M=1$ 時

- 一次多項式畫出來是一條可以有斜率的直線。
- w_0 大約是0.8， w_1 大約是-1.2，可以產出最小的 $E(w)$ 值。

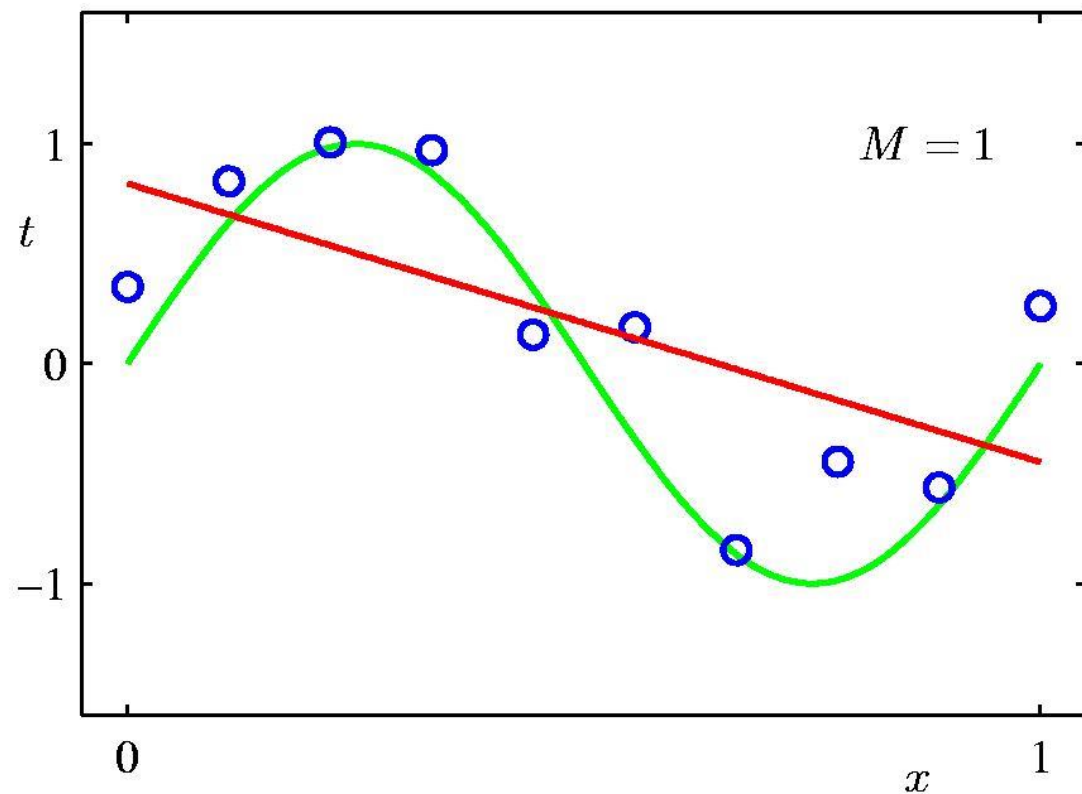


Figure 1.4

當 $M=3$ 時

- 看來和10個資料點貼合的不錯。
- 作者有把綠線畫出來，所以看起來三次是個好選擇，但在實際資料集時，我們並不知道資料點是由什麼模型產生的，我們只能假設我們演算法的模型，然後用資料集去測。

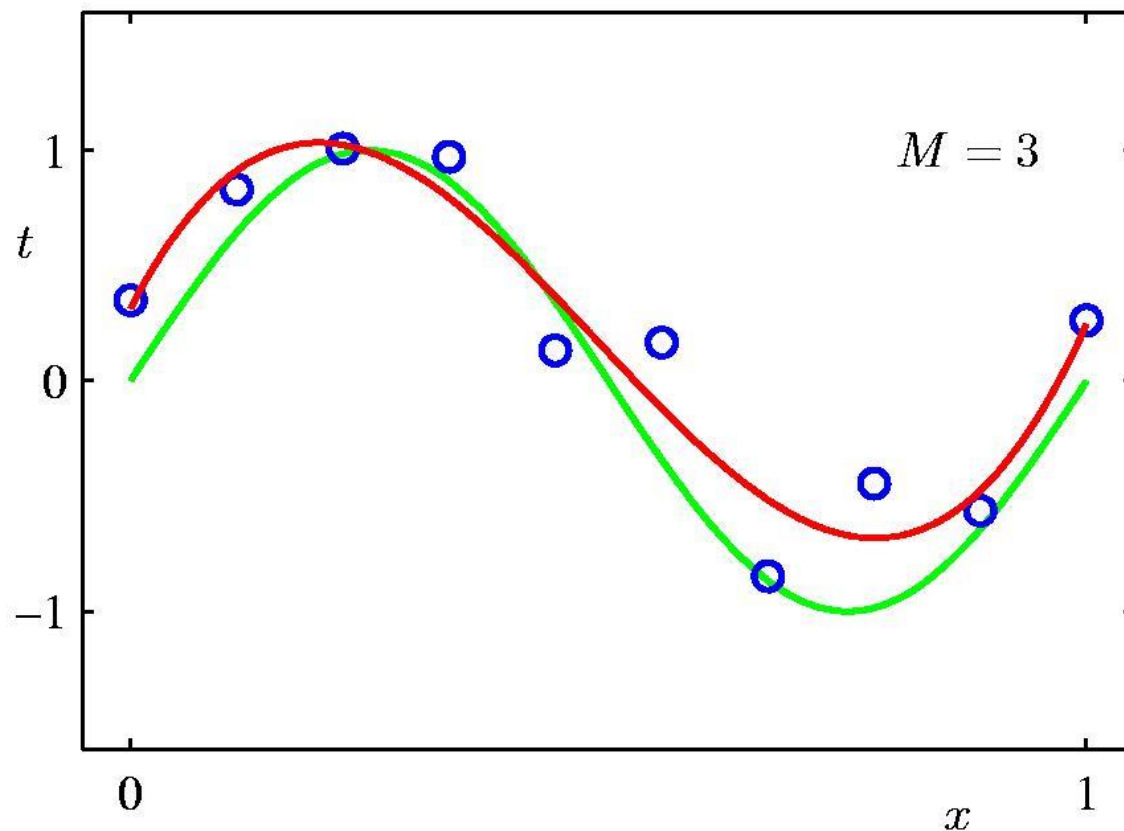


Figure 1.4

當 $M=9$ 時

- 因為你知道真正的資料是由綠線產生出來的，所以九次多項式離譜。
- 但九次多項式的錯誤值是0，因為每一個都穿過了，沒有任何一個點有偏離。

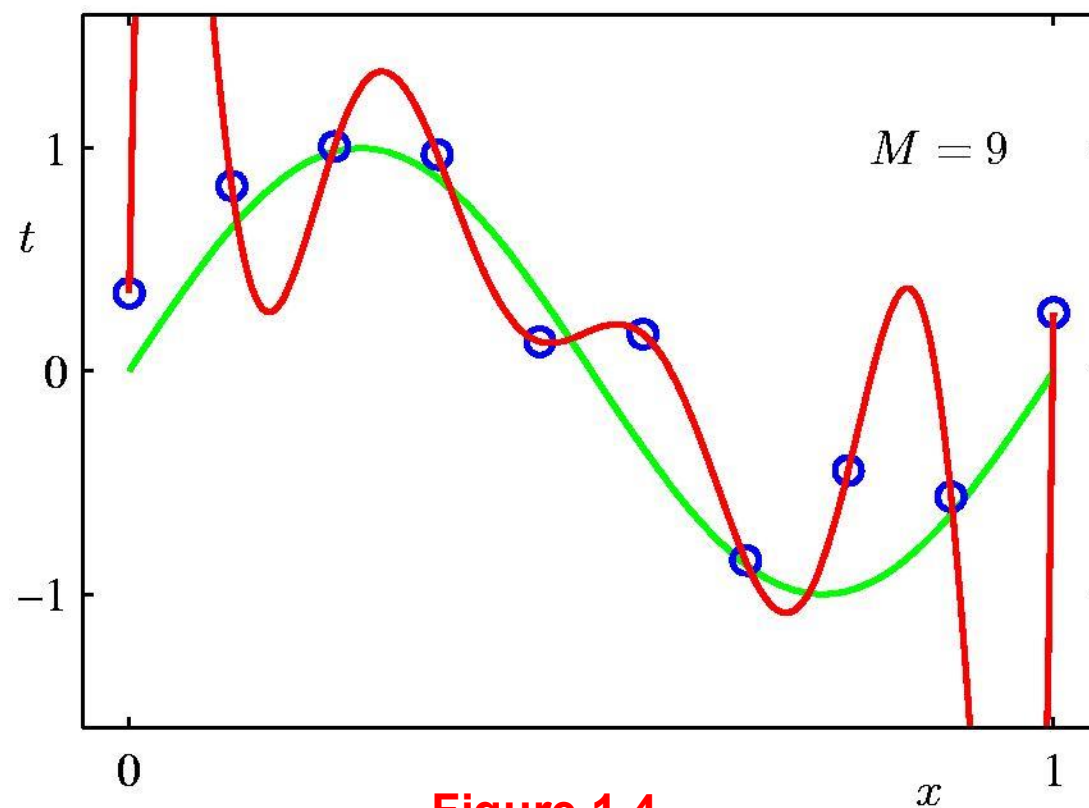


Figure 1.4

過度擬合

- 模型太過複雜，參數太多，造成為了最小化訓練資料的錯誤，權重太貼合訓練資料，而脫離了真實的來源模型。
- 而測試資料是從真實的來源模型來，但我們的演算法沒有看過，就會產生了很大的錯誤。

Root-Mean-Square 均方根

w^* 最佳參數組

$$E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^*)/N} \quad (1.2)$$

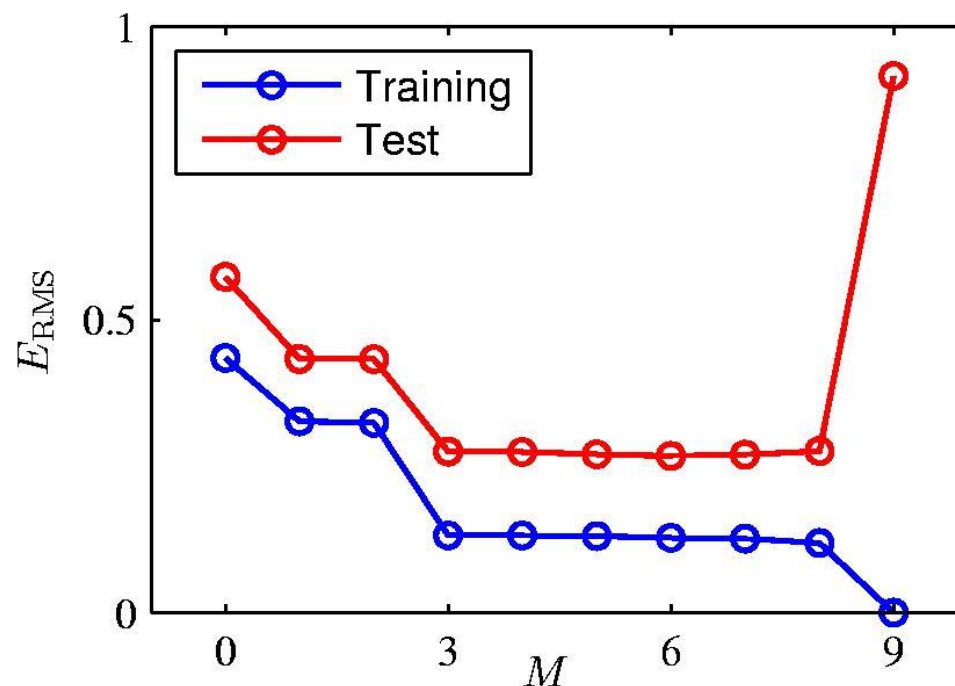


Figure 1.5

最小錯誤時，權重(多項式係數)的值

	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

數值這麼大，不正常

Table 1.1

如果我們的資料量不是10筆而是15筆呢？

- 九次多項式不再出現過度擬合的狀況，預測結果不錯。

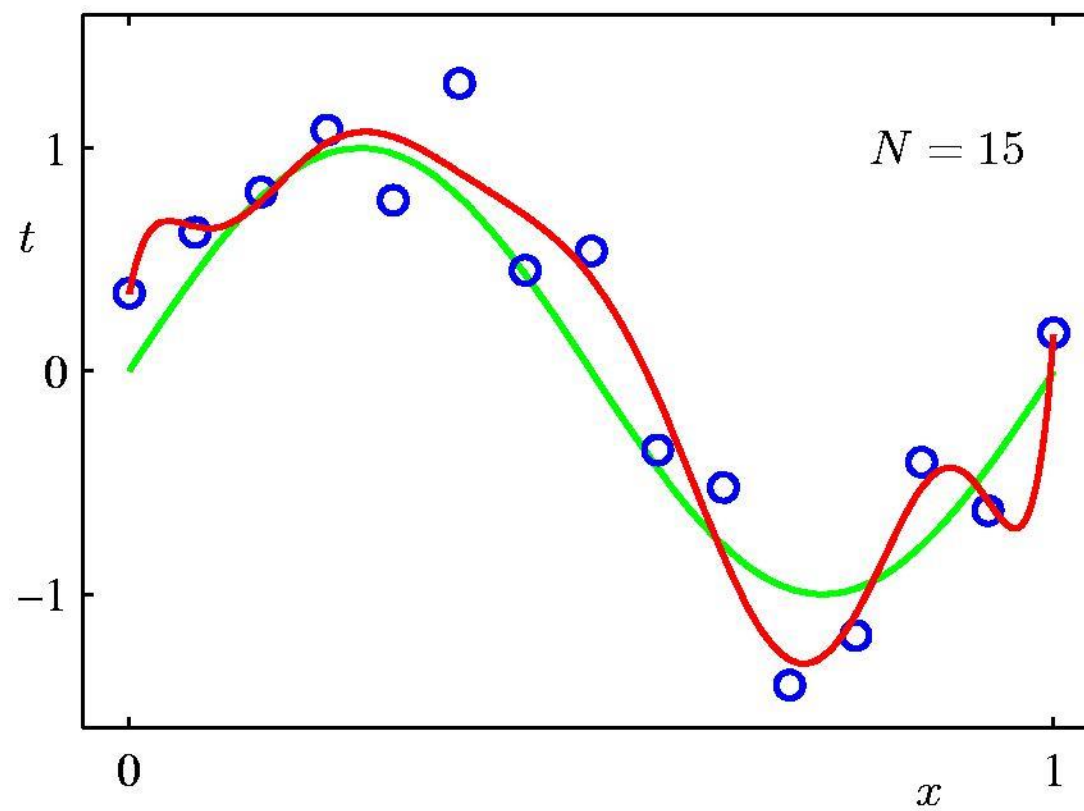


Figure 1.5

如果是100筆呢？

- 九次多項式沒有過度擬合，模型表現良好。
- 對於10筆資料的小資料集，九次多項式有10個參數，自由度太高了，可以緊緊地貼合少少的資料，發生了過度擬合。
- 對於100筆資料的大資料集，10個參數的自由度只能為了讓錯誤值變小而彼此妥協，反而是抵消了雜訊。

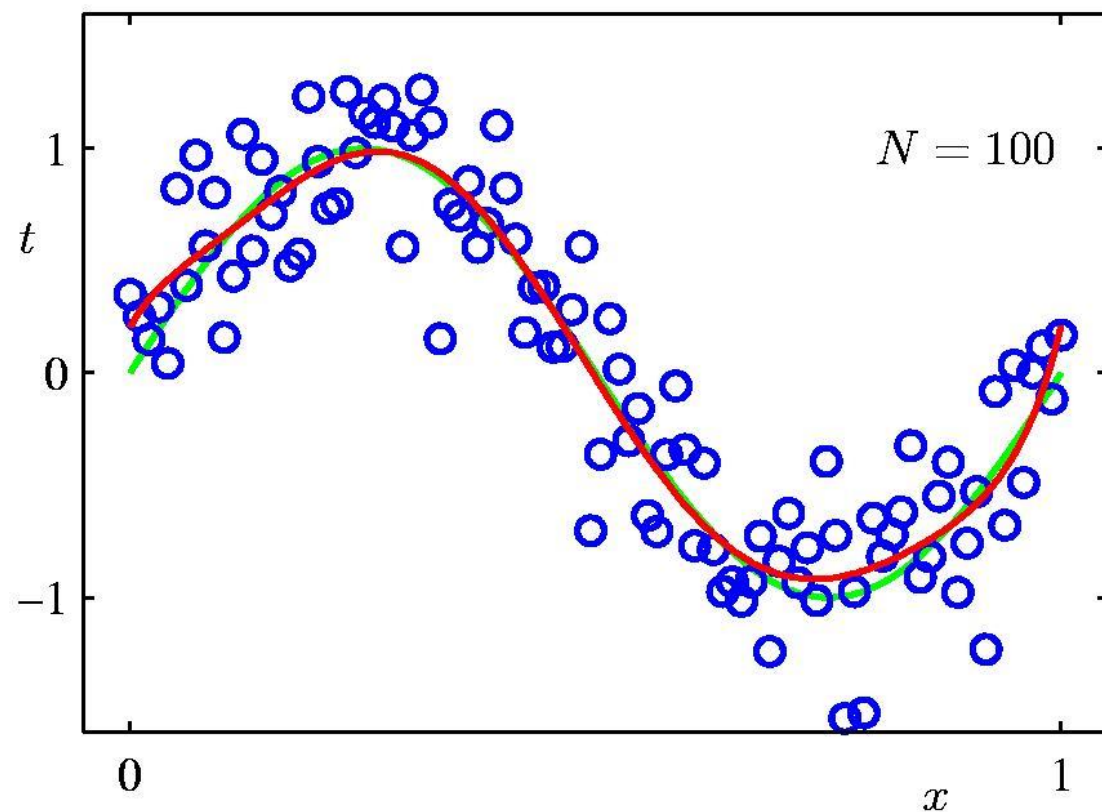


Figure 1.5

正規化(或稱正則化)

- 如果我們覺得超大係數很怪，那我們可以壓抑它。將係數的大小也算成錯誤的一部份。
- 右式紅色的部份在數學上稱為正則項，其實也是個平方和。
- 兩種平方和要怎麼平衡？我們就引入了另一個超參數 λ 。 λ 愈大，表示我們愈厭惡大的係數。 λ 愈小，表示愈不在乎。

$$\tilde{E}(\mathbf{w}) = \underbrace{\frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2}_{\text{殘差平方和}} + \underbrace{\frac{\lambda}{2} \|\mathbf{w}\|^2}_{\text{正則項}} \quad (1.4)$$

$$\|\mathbf{w}\| = \sqrt{w_0^2 + w_1^2 + \cdots + w_M^2}$$

當 $\lambda = e^{-18} \sim 0.0000000152$

- 這仍然是九次多項式，但加入了正則項。
- 就算 λ 很小，係數仍然被被壓抑，避免了過度擬合的狀況。

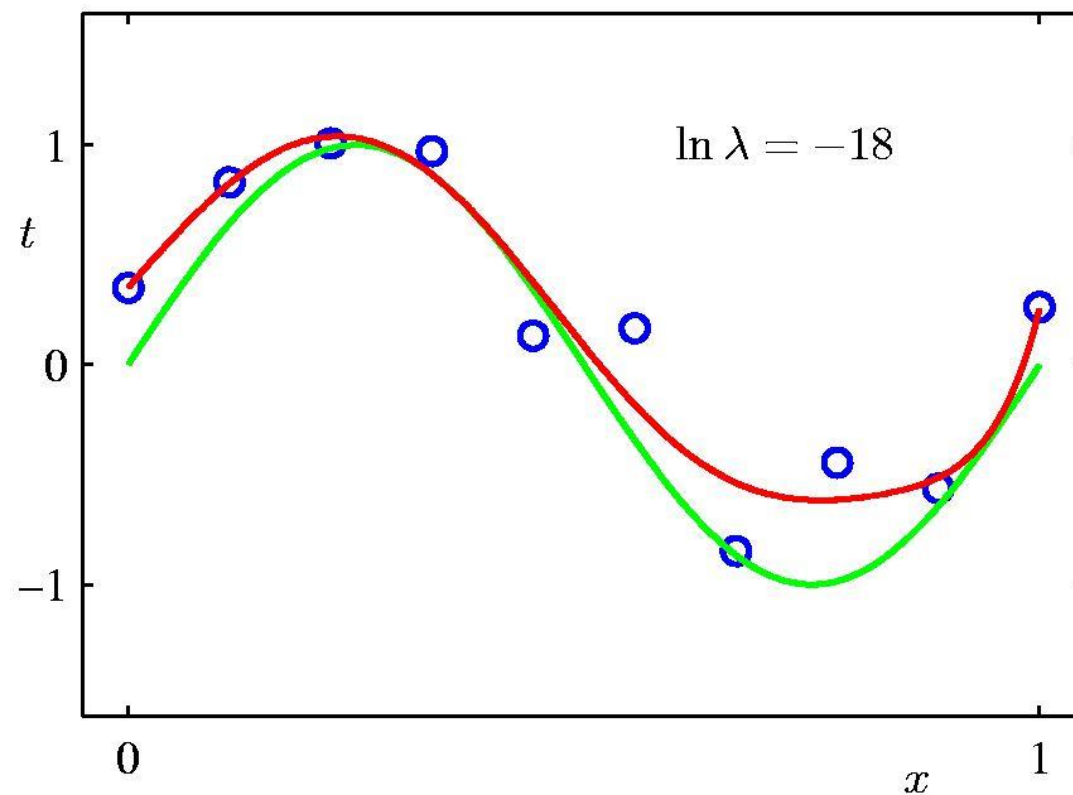


Figure 1.7

當 $\lambda = e^0 = 2.718$

- 多項式係數被壓得很小，九次多項式看起來像零次多項式。

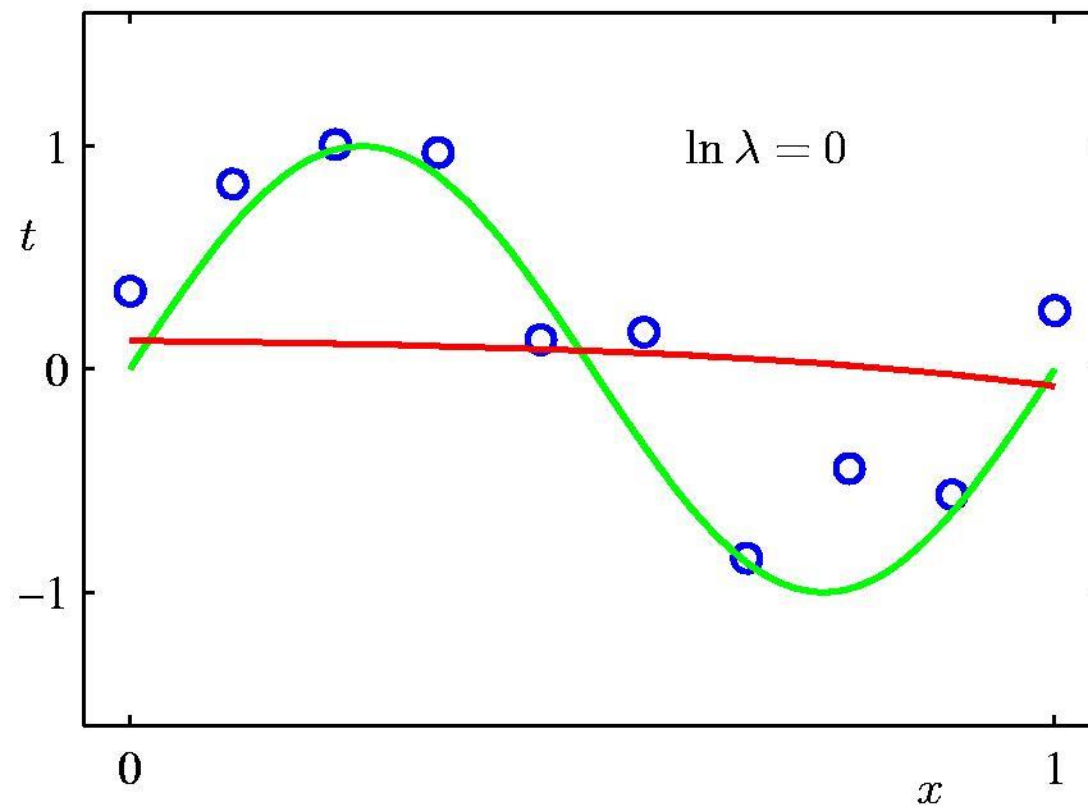


Figure 1.7

調 λ 這個超參數

- λ 非常接近0時，就像沒加，所以有明顯過擬合的現象， λ 到 e^{-35} 左右，過擬合的現象就被不明顯了。
- 所謂過擬合現象：訓練資料集的錯誤值很小，但測試資料集的錯誤值很大。
- 當 λ 大於 e^{-25} 時，模型太注重正則項，而讓殘差項變大，這個九次多項式愈來愈對不準訓練資料集。
- e^{-35} 到 e^{-25} 之間稱為甜蜜區間，是模型表現最好的時候。

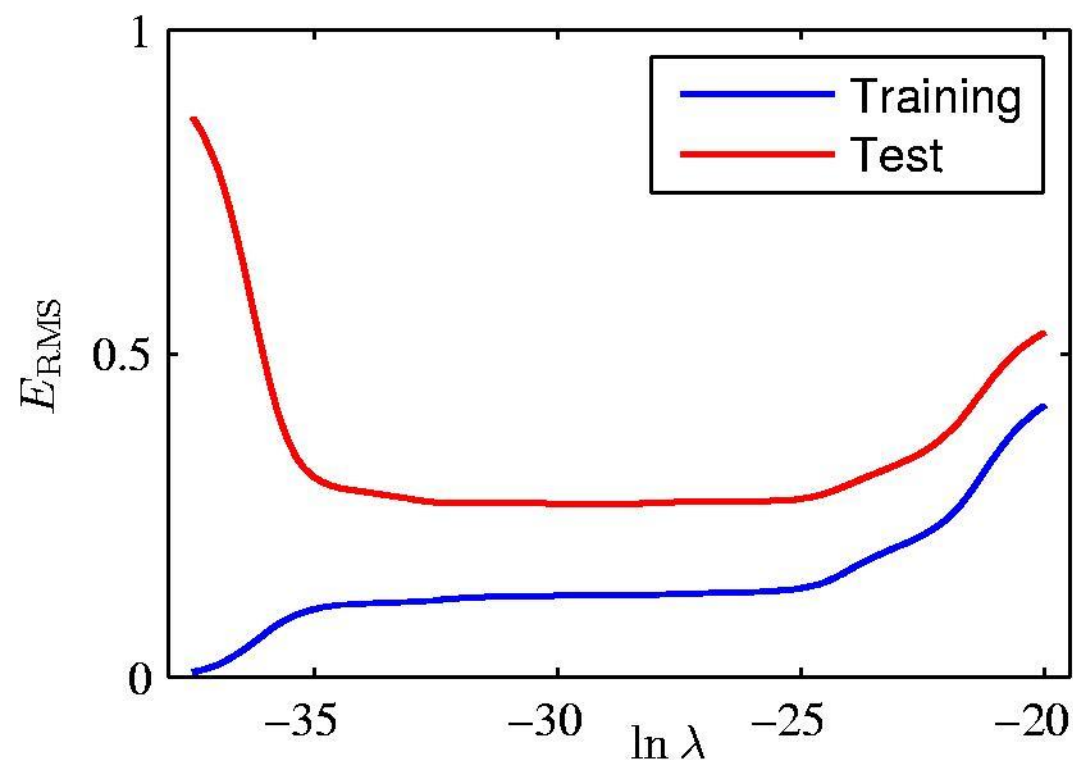


Figure 1.8

看數值

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
w_0^*	0.35	0.35	0.13
w_1^*	232.37	4.74	-0.05
w_2^*	-5321.83	-0.77	-0.06
w_3^*	48568.31	-31.97	-0.05
w_4^*	-231639.30	-3.89	-0.03
w_5^*	640042.26	55.28	-0.02
w_6^*	-1061800.52	41.32	-0.01
w_7^*	1042400.18	-45.95	-0.00
w_8^*	-557682.99	-91.53	0.00
w_9^*	125201.43	72.68	0.01

Table 1.2

不夠

較好

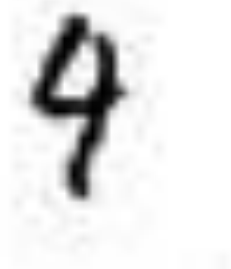
過頭

調參數

- 超參數會強烈地影響模型的表現，資料不同，甜蜜區間的範圍也不一樣，所以調參數是機器學習裏得常作的事，這裏只有一個超參數 λ ，而且它的影響是連續的，可以很容易看得出來甜蜜區間在哪裏。
- 當你有十個超參數，而且影響是不是連續的你不知道，要暴力跑完所有超參數的組合會用掉一兩個禮拜時，那調參數就是藝術了。

機率論

- 課本裏會有很多機率的東西，因為機器學習本來就是要算機率的工作。
- 例如說MNIST裏的這張圖



- 它到底是4還是9？

- LeNet算的其實是機率。不只是4和9，它算了0到9總共十個機率值，以機率最大的那個當作它的答案。
- 在羅吉斯迴歸時，你們看到我們很明顯地在算機率。因為羅吉斯函數就是一個將任意實數映射到 $[0,1]$ 之間的函數，那個 $[0,1]$ 就是機率值的範圍。

難處

- 再者，資料一定會帶著雜訊。例如說鳶尾花資料集，是野地收集來的，手量會有雜訊，只記到小數點第1位，消失的位數也是雜訊。
- 要模擬雜訊，我們通常會靠機率模型，因為雜訊本來就是每次出來的結果都不一定，這就是機率模型的特性。
- 但你們二下才會學機率與統計，而我們現在需要用到。所以我會提，但不太講得太深，只到讓你們知道個概念而已。

課本上的例子

- 機率論要處理的是不確定的因果關係。如果我們有兩個袋子，一個紅色，一個藍色，裏面各有一些橘子和青蘋果。
- 如果我們選到紅袋子的機率是0.4，選到藍袋子的機率是0.6，那我們最後會拿出青蘋果的機率是多少？

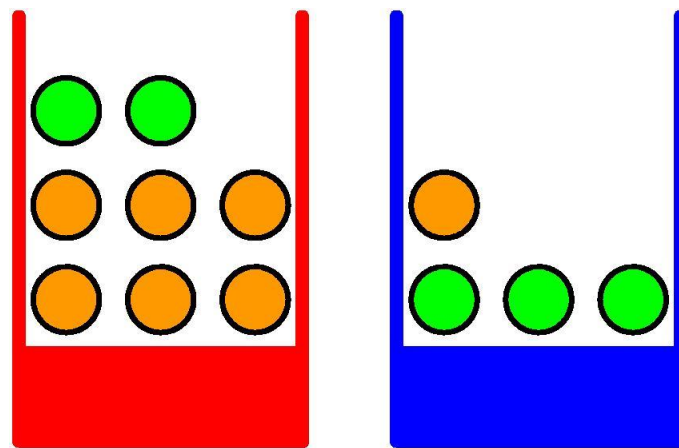


Figure 1.9

回想高中的機率課

總機率

=選到紅袋子的機率*紅袋子裏取到蘋果的機率+選到藍袋子的機率
*藍袋子裏取到蘋果的機率

$$= 0.4 \times \frac{2}{8} + 0.6 \times \frac{3}{4}$$

$$= 0.1 + 0.45$$

$$= 0.55$$

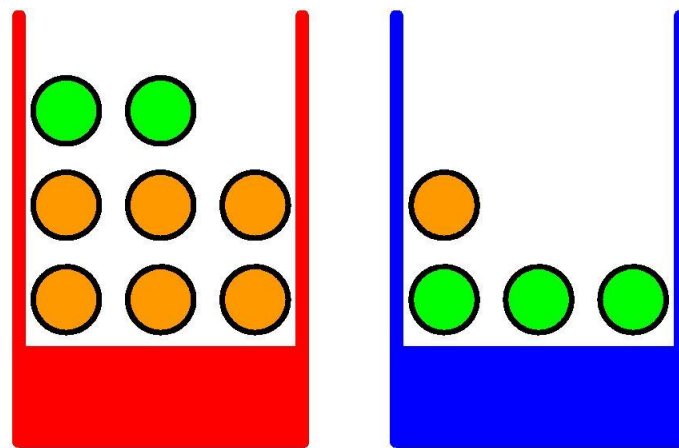
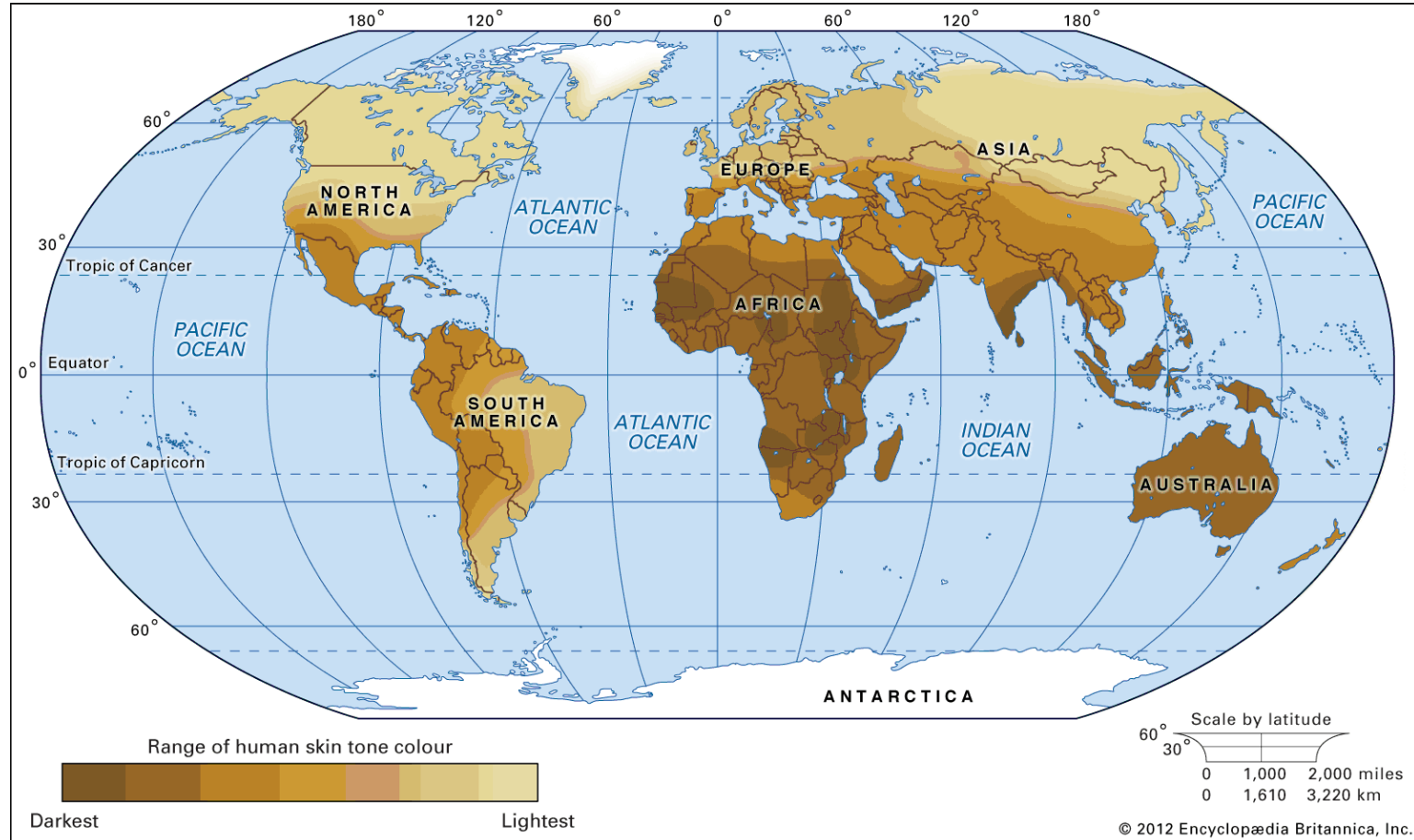


Figure 1.9

第二個問題

- 如果有人拿出一個橘子，那這個橘子是從藍袋子裏拿出來的機率是多少？
- 貝氏定理問題，我們講完貝氏定理之後解。
- 我去問Gemini 2.5 flash「如果我們有兩個袋子，一個紅色，一個藍色，紅袋子裏6個橘子和2個蘋果，藍袋子裏有1個橘子3個蘋果。假設我們選到紅袋子的機率是0.4，選到藍袋子的機率是0.6，如果有人拿出一個橘子，那這個橘子是從藍袋子裏拿出來的機率是多少？」它給我完全正確的解答，而且還有貝氏定理的說明。

世界各洲人口膚色分佈圖



名詞解釋：聯合機率、邊際機率、條件機率

- 參加奧運的選手，來自五大洲，所以我們訂「洲別」是第一個事件。人的膚色，大致分為「白皙、中等、深色」，我們訂為第二個事件。
- 我們先看兩個事件就好，比較單純。
- 按照洲別和膚色，我們可以畫出一個統計表

					c_i 第i欄的總數
y_j			n_{ij}		r_j 第j列的總數
					x_i

Figure 1.10

名詞解釋：聯合機率、邊際機率、條件機率

- 聯合機率：格裏的數除總數

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \quad (1.5)$$

- 邊際機率：只考慮一種事件

$$p(X = x_i) = \frac{c_i}{N}. \quad (1.6)$$

- 條件機率：

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i} \quad (1.8)$$

- 例如：來自亞洲，而且是中等膚色的機率是多少？

- 例如：來自亞洲的機率是多少？

- 例如：在所有亞洲選手中，中等膚色的機率是多少？

名詞解釋：加法律、乘法律

- 邊際機率是怎麼來的？就是同一個事件的聯合機率加起來的。這就是加法律。
- 聯合機率等於邊際機率乘條件機率，這就是乘法律。

$$\begin{aligned} p(X = x_i) &= \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^L n_{ij} \\ &= \sum_{j=1}^L p(X = x_i, Y = y_j) \quad (1.7) \end{aligned}$$

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned} \quad (1.9)$$

來自亞洲，而且是中等膚色的機率
= 來自亞洲的機率 x 在所有亞洲選手中，中等膚色的機率

課本上的例子

有兩個變數 X 和 Y
 X 的範圍被劃分成9個bin
 Y 的值只有1和2兩種可能
抽60個點，分佈如左上

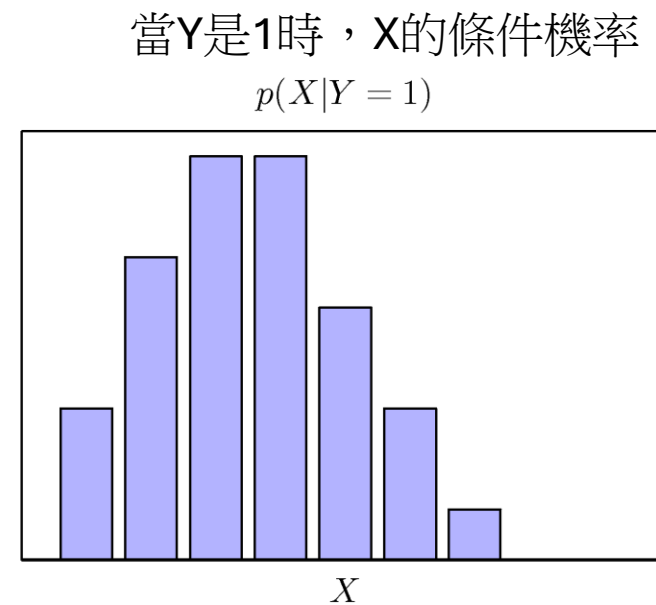
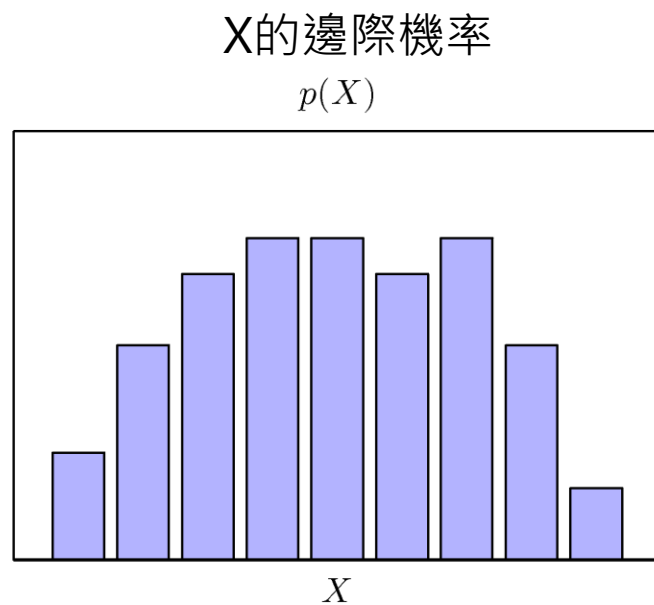
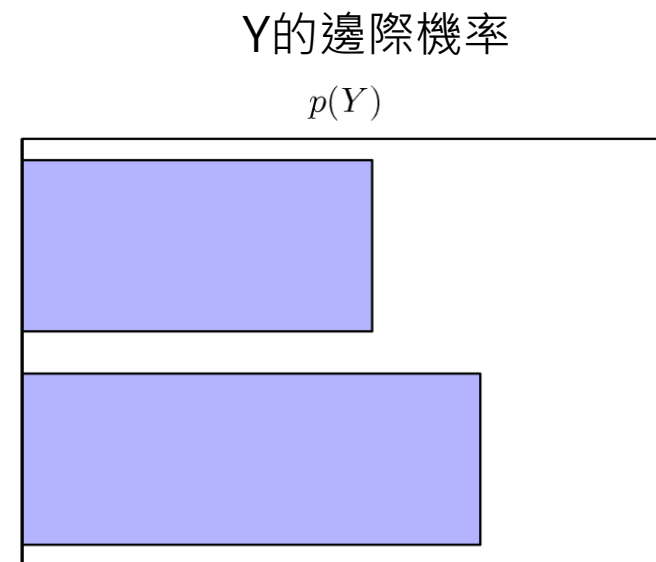
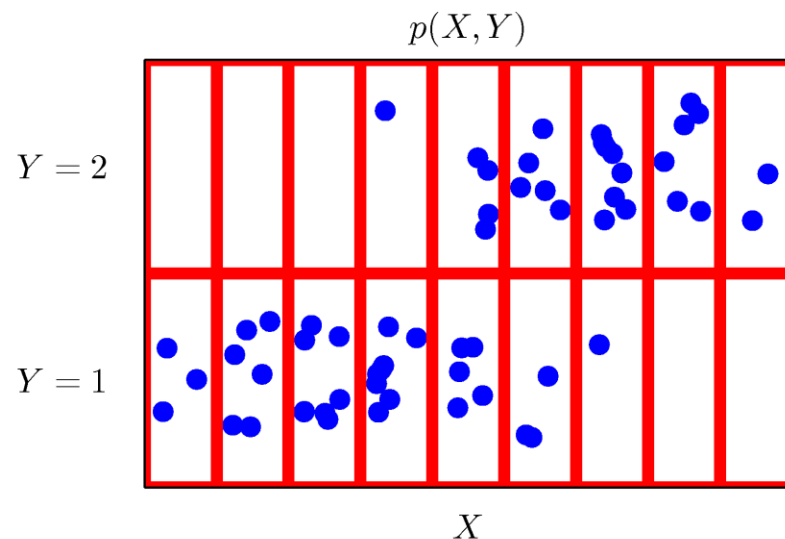


Figure 1.11

貝氏定理

- 它是一個描述機率關係的數學公式，當我們獲得新的證據或是資訊時，如何更新一個事件的機率。
- 回到剛剛的奧運會的例子，如果你不知道選手001號來自哪一洲，你要猜他的膚色只能用世界人口比例，白15%，中等45%，深色40%。
- 當你知道他是來自非洲時，因為非洲人中深膚色的比例很高，那你猜他的膚色機率就會把深色皮膚推高。

貝氏定理的數學式

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \quad (1.12)$$

- X和Y是兩種不同的事件。
- 原因：聯合機率有兩種不同的乘法律表示法，移項就會得到貝氏定理了

$$p(X, Y) = p(Y|X)p(X)$$

$$p(X, Y) = p(X|Y)p(Y)$$

貝氏定理的口號

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \quad (1.12)$$

- 後驗機率 正比於 概似性 乘 先驗機率 (1.44)
- 後驗機率指的是 $p(Y|X)$ ，是我們最終想知道的的機率，在已經觀察到X事件後，Y事件的機率。
- 概似性指的是 $p(X|Y)$ ，是已經Y已經發生的狀況下，觀察到X事件的機率。
- 先驗機率指的是 $p(Y)$ ，這是我們還沒觀察到X事件時，對Y事件發生機率的初始估計。
- 為什麼他們說"正比"呢？因為這個口號把 $p(X)$ 忽略了，只剩 $p(Y|X)$ 、 $p(X|Y)$ 、 $p(Y)$ ，所以等式不成立，後驗機率不等於概似性乘先驗機率，但因為已經觀察到X時，就像我們已經知道某位選手是亞洲人， $p(X)$ 就固定了，不論Y膚色是哪一種，都不會影響 $p(X)$ 。

貝氏定理跟機器學習的關係

- 機器學習的任務：從數據中學出一個模型，然後用這個模型來預測或分類。
- 當我們假定一個模型，例如說一個三次多項式，我們就會有四個權重，就是三次多項式的四個參數，我們下一步就是問，這給了這四個參數值，產出我們的數據的機率有多高？
- 最好的權重，就是能讓產出我們訓練數據機率最高的那一組。
- 用數學符號表示

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \quad (1.43)$$

續

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

- \mathbf{w} 指的是你模型的權重值
- \mathcal{D} 指的是你的觀察數據
- 什麼是最好的權重值？就是能讓後驗機率最大的權重值。

貝氏定理練習：回到剛剛橘子與蘋果的問題

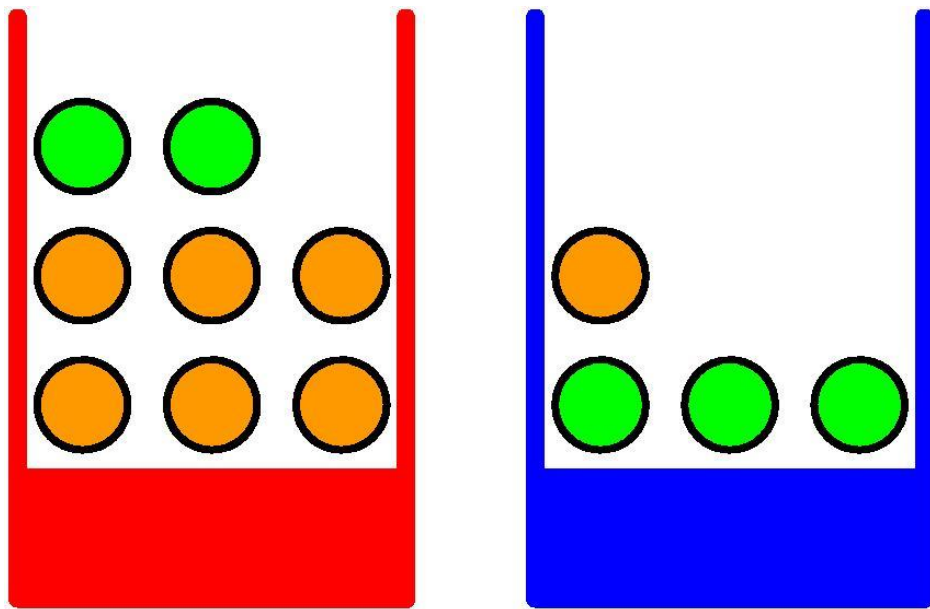


Figure 1.9

- 選到紅袋子的機率是0.4，選到藍袋子的機率是0.6，如果有人拿到的是一個橘子，這個橘子從藍袋子裏拿出來的機率是多少？

聯合機率有兩種不同的乘法律表示法

- 設X是取出的水果，Y是袋子的顏色

聯合機率值表

	$X = \text{orange}$	$X = \text{apple}$
$Y = \text{blue}$	$0.6 * 1/4 = 0.15$	$0.6 * 3/4 = 0.45$
$Y = \text{red}$	$0.4 * 6/8 = 0.3$	$0.4 * 2/8 = 0.1$

$$p(Y = \text{blue} | X = \text{orange}) = \frac{p(X = \text{orange}, Y = \text{blue})}{p(X = \text{orange})} = \frac{0.15}{0.45} = 0.3333$$

機率密度

- 水果種類、洲別都是離散型的資料。
- 若資料是連續型的，像是長度、重量、降雨量、風速、時間，機率值是要先訂好一個區間才有值，它是一個積分的結果，所以我們說它是機率密度。
- 例如說，這是一個合理的問法：台灣人身高介於170.0與171.0公分的機率是多少。
- 如果問：台灣人身高為170.312495公分的機率是多少？沒有什麼意義，這其實是一個密度，要積分後才會得到機率。

視覺化

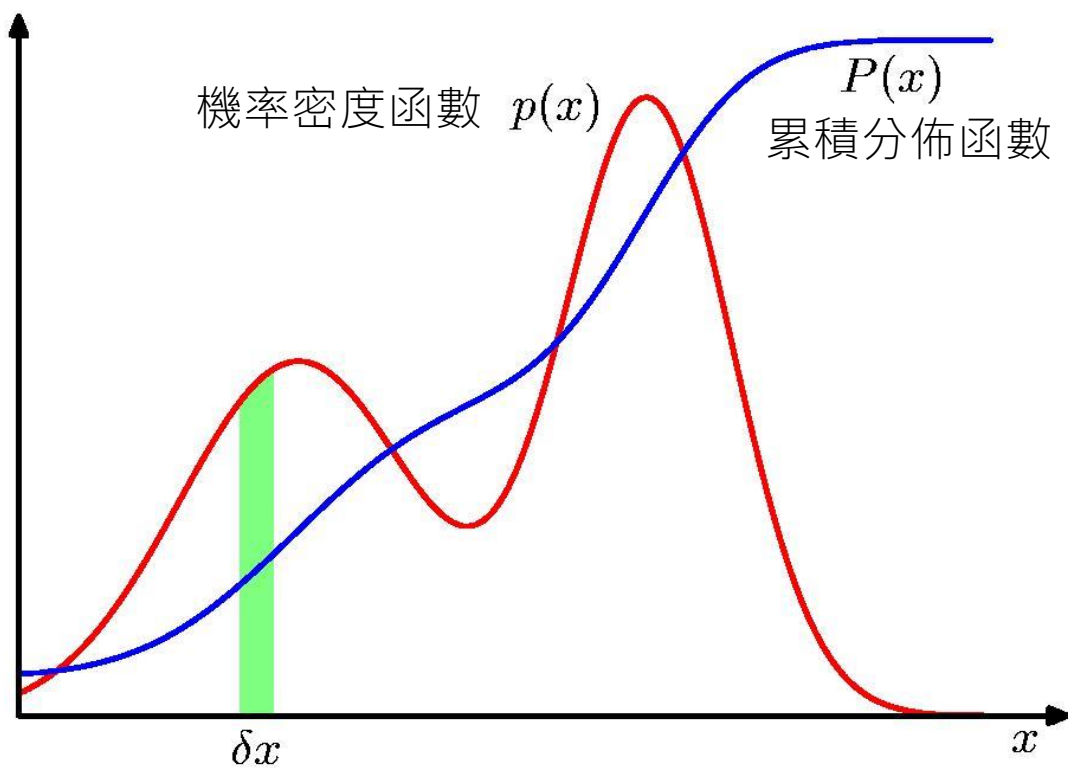


Figure 1.2

- 綠色區域的面積，指的是機率。
- 這個面積是對機率密度函數積分得出來的

$$p(x \in (a, b)) = \int_a^b p(x) dx \quad (1.24)$$

- 因為機率的總和為1

$$\int_{-\infty}^{\infty} p(x) dx = 1 \quad (1.26)$$

累積分佈函數

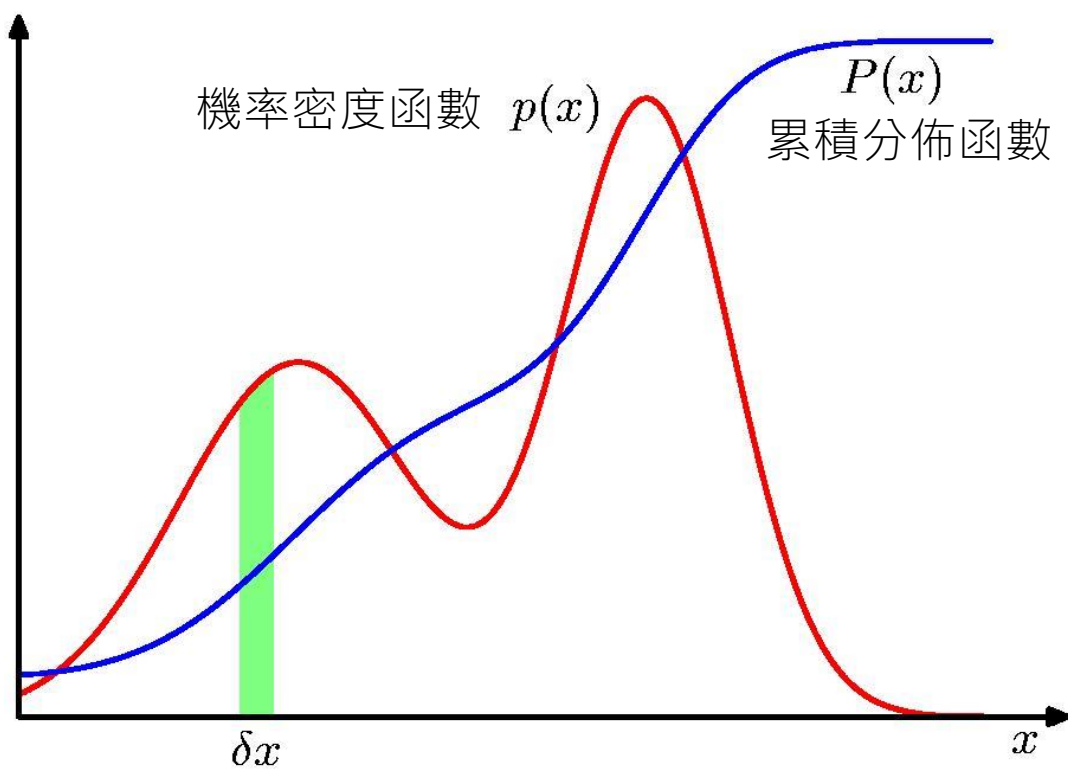


Figure 1.2

- 就是這個積分一直積到 z 值為止
$$P(z) = \int_{-\infty}^z p(x) dx \quad (1.28)$$
- 所以它一定是漸增函數，因為機率密度都是大於等於0
- 一個比喻，就是台灣人身高小於 x 公分的機率值。

期望值

- 它僅是平均而已。
- 如果你有台灣人的身高分佈機率密度函數，你可以用來求平均身高。

$$\mathbb{E}[f] = \int p(x)f(x) \, dx \quad (1.34)$$

- 其中的 x 是身高， $f(x)$ 是我們想求的目標值，在這個例子中 $f(x) = x$ ，如果你的目標值是身高的平方，則 $f(x) = x^2$
- 如果我們有性別的條件，或是年紀的條件，或是居住縣市的條件

$$\mathbb{E}_x[f|y] = \int p(x|y)f(x) \quad (1.35)$$

離散型

- 例如 x 是家戶子女數，當 $f(x) = x$ 時，台灣平均家戶子女數為

$$\mathbb{E}[f] = \sum_x p(x)f(x)$$

- 條件期望值 $\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x)$

- 這條件可能是年紀、居住地、家戶收入區間、種族

- 當離散型的抽樣量愈大時，愈接近連續型 $\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$

變異數和共變異數

$$\text{var}[f] = \mathbb{E} \left[(f(x) - \mathbb{E}[f(x)])^2 \right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2 \quad (1.39)$$

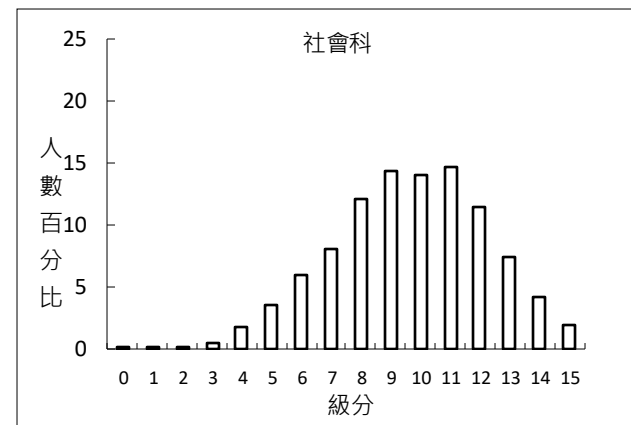
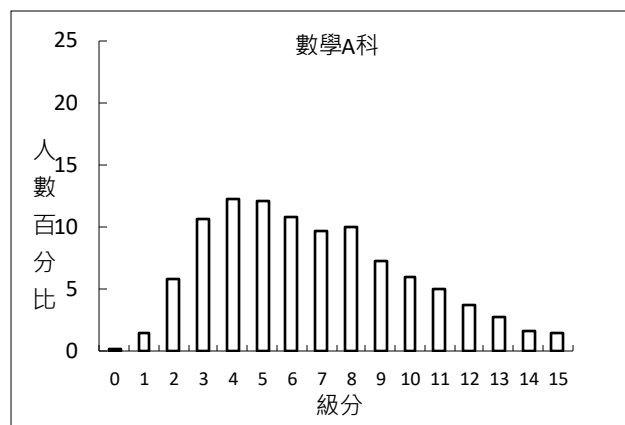
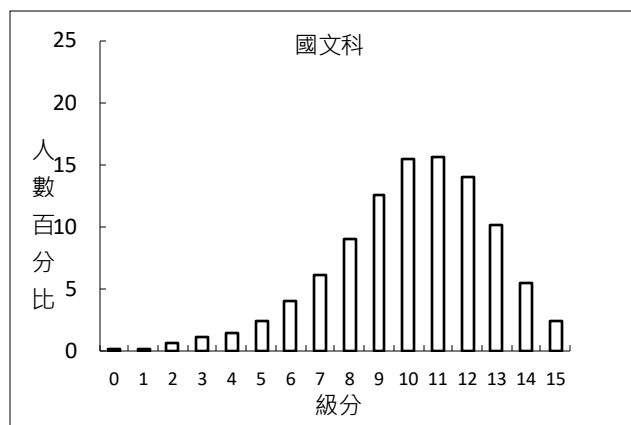
$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{x,y} [\{x - \mathbb{E}[x]\} \{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y} [xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \quad (1.41)$$

當 \mathbf{x}, \mathbf{y} 是向量
時，如此表示

$$\begin{aligned} \text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x}, \mathbf{y}} [\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\} \{\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T]\}] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}} [\mathbf{x} \mathbf{y}^T] - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}^T] \end{aligned} \quad (1.42)$$

高斯分佈

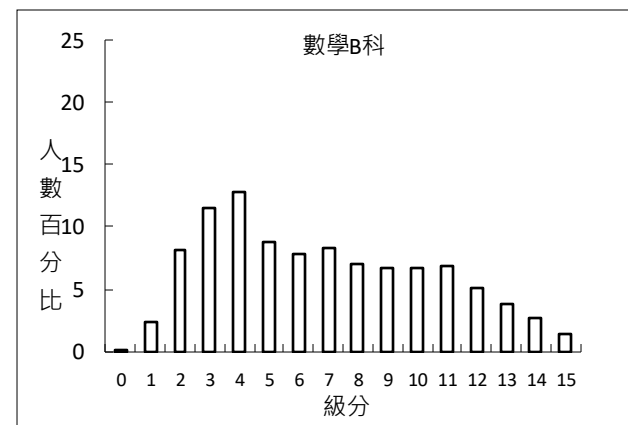
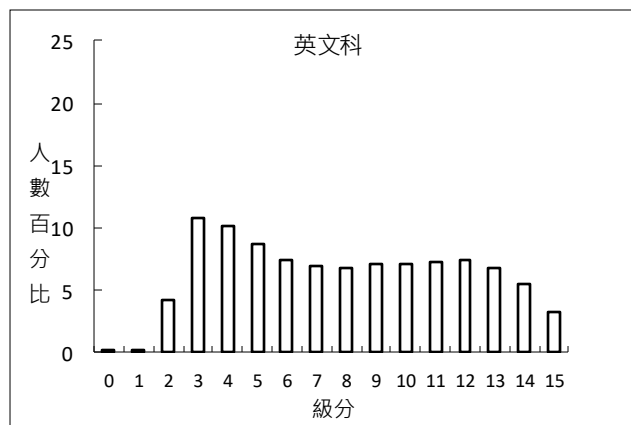
- 又稱常態分佈，是你們最常遇到的連續型機率分佈，身高、體重、考試成績、螺絲釘長度、收入、日均溫，用高斯分佈來建模，效果常常很好。



114學年學測成績

不吻合的時候

- 英文的成績是雙峰分佈，不會是高斯，因為高斯是單一高峰。
- 數學B左右不對稱，也不會是高斯，因為高斯左右對稱。



高斯分佈的機率密度函數

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \quad (1.46)$$

- 它有兩個參數 μ, σ ，其義意是平均值和標準差。
- 畫出來的圖形，有一個高峰在中間平均值處，離平均值愈遠，機率密度愈低，而且對稱。

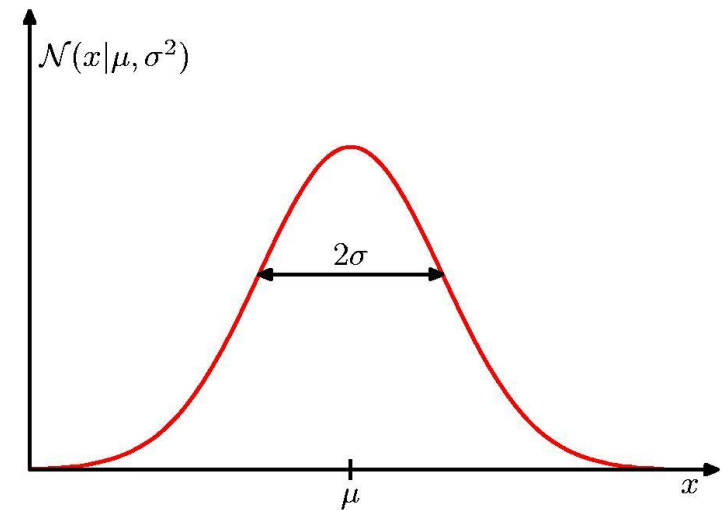
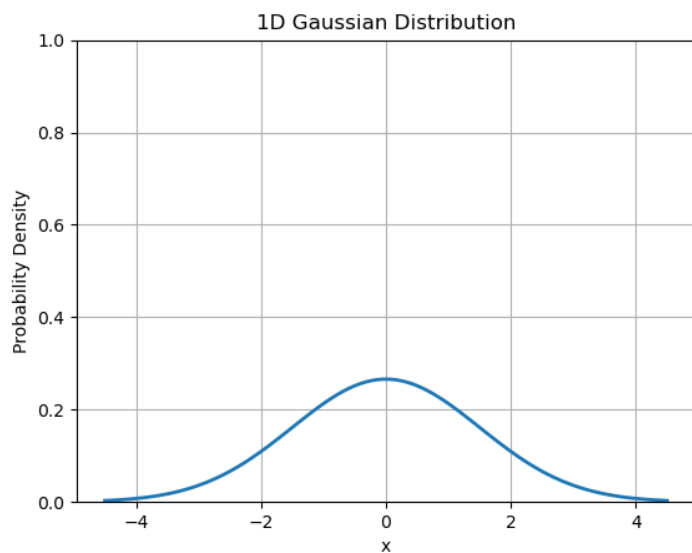
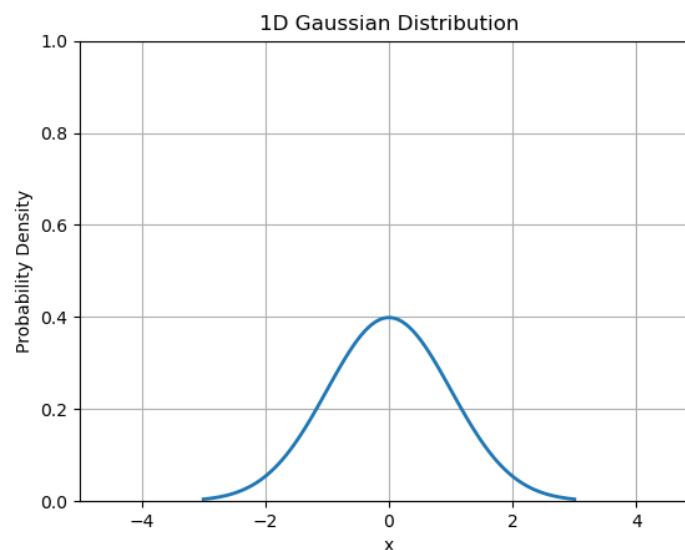


Figure 1.13

標準差愈小，
密度愈集中



1.5



1.0

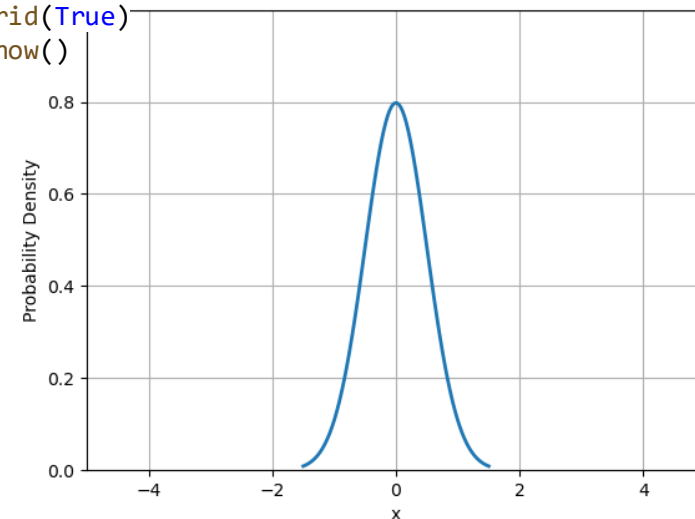
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# Define the mean and standard deviation
mu = 0
sigma = 0.5

# Create a range of x values
x = np.linspace(mu - 3*sigma, mu + 3*sigma, 100)

# Calculate the probability density function (PDF)
y = norm.pdf(x, mu, sigma)

# Plot the Gaussian distribution
plt.plot(x, y, linewidth=2)
plt.xlabel('x')
plt.ylabel('Probability Density')
plt.title('1D Gaussian Distribution')
plt.grid(True)
plt.show()
```



0.5

平均值和變異數

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \, dx = \mu \quad (1.49)$$

- 平均值蠻直覺的，因為高斯左右對稱，均值一定在對稱軸上，高斯的對稱軸又在 $x = \mu$ 處，所以均值就會是 μ 。

$$\text{var}[x] = \mathbb{E}[(x - \mathbb{E}[x])^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) (x - \mu)^2 \, dx = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2 \quad (1.51)$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \, dx = \mu^2 + \sigma^2 \quad (1.50)$$

證明平均值為 μ (練習1.8)

由定義 $\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x dx$

得到 $\mathbb{E}[x] = \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} x dx$

變數變換 $y = x - \mu$

得到 $\mathbb{E}[x] = \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} y^2 \right\} (y + \mu) dy$

$$= \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} y^2 \right\} y dy + \int_{-\infty}^{\infty} \left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left\{ -\frac{1}{2\sigma^2} y^2 \right\} \mu dy$$

此項為0，因為 y 和 $-y$ 互消

此積分與 μ 無關，是 y 的機率密度函數之和，為1

$$= \mu$$

證明變異數為 σ^2

由機率密度
函數的定義

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\}$$

和其積分為1
的性質

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$$

移項可得

$$\int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} dx = (2\pi\sigma^2)^{1/2}$$

等式兩邊都
對 σ^2 微分

$$\int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} \left(-\frac{1}{2} \cdot -1 \cdot \frac{1}{\sigma^4}(x - \mu)^2 \right) dx = (2\pi)^{1/2} \frac{1}{\sqrt{\sigma^2}} \frac{1}{2}$$

消去1/2，移
項 $(2\pi)^{1/2}$ 兩
邊都乘上 σ^3

$$(2\pi\sigma^2)^{-1/2} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} (x - \mu)^2 dx = \sigma^2$$

係數移入積
分項，即為
變異數定義

$$\int_{-\infty}^{\infty} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\} (x - \mu)^2 dx = \sigma^2$$

利用LLM給你證明

- 我用Gemini 2.5 flash，一步步的數學式都正確。

Gemini

2.5 Flash ▾

證明給我，高斯分佈的變異數為sigma平方

- ◆ 首先，高斯分佈 (Gaussian distribution) 的機率密度函數 (probability density function) 為：

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

其中 μ 是平均數 (mean)， σ^2 是變異數 (variance)。

變異數的定義為：

$$\text{Var}(X) = E[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

將高斯分佈的機率密度函數代入變異數的公式：

$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

步驟一：變數代換

問問 Gemini



Deep Research



圖像



Canvas



引導式學習



Gemini 可能會出錯，請查證。

雙變數高斯分佈

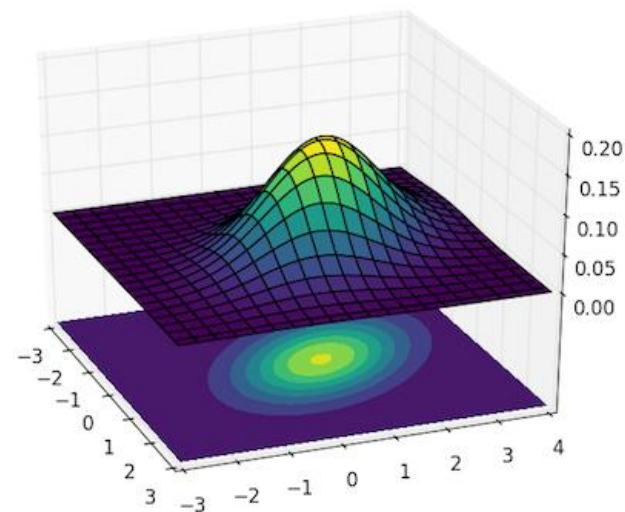
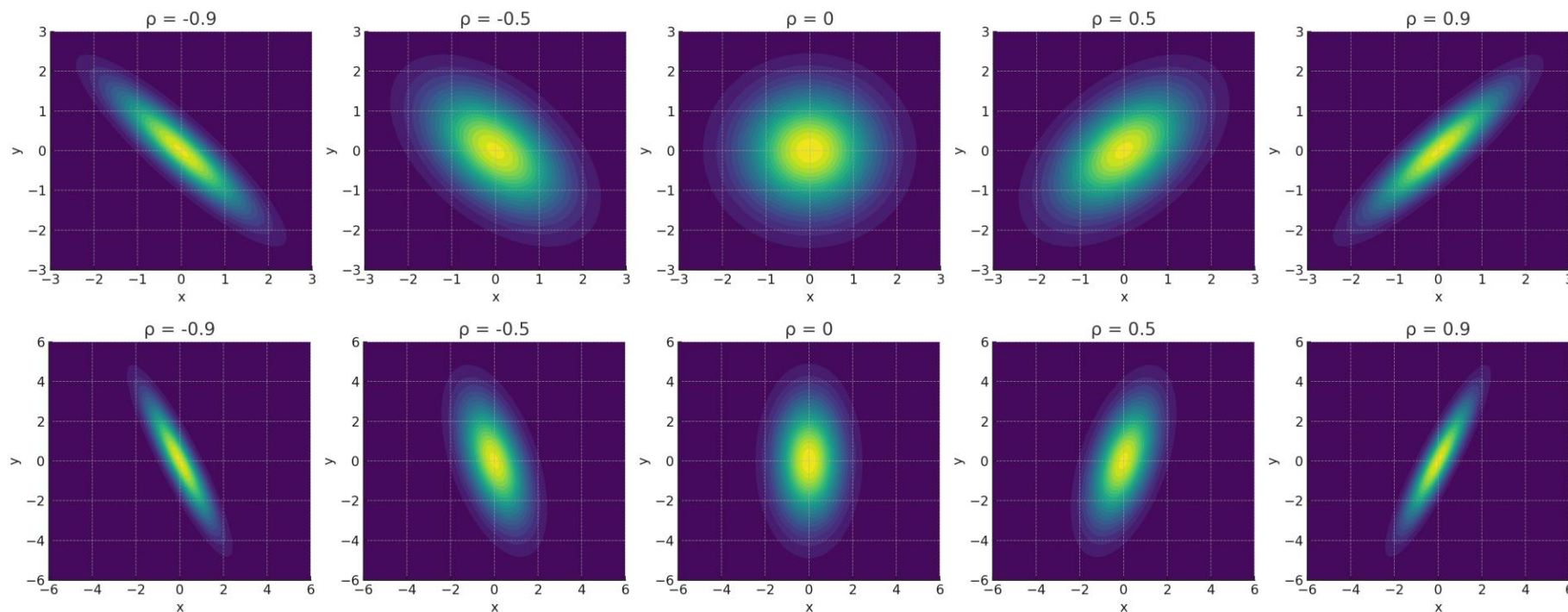
- 當變數不只一維時，還是有高斯分佈。
- 雙變數的例子

$$f(x, y; \mu_x, \mu_y, \sigma_x, \sigma_y, \rho) \\ = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2} \right]\right)$$

- 其中 ρ (rho)是變數 x 和 y 之間的皮爾森相關係數，其值介於 -1 到 1 之間，它衡量了兩個變數之間的線性相關性。

視覺化

- 比較亮的部分，機率密度比較大
- $\mu_x = \mu_y = 0$ 上排 $\sigma_x = \sigma_y = 1$ 下排 $\sigma_x = 1, \sigma_y = 2$



生活中的實例

- 身體與體重。
- 各別都是常態分佈，但兩者之間有相關性，比較高的人重的機率較大。
- 國文和數學A的學測成績。
- 各別都是常態分佈，但兩者之間有相關性，用功的學生兩科的分數都在平均之上，落後的學生，兩科都在平均之下。

多變數高斯分佈

- 當我們要同時考量身高、體重、和數學A的成績，就會有三個變數。多個變數的高斯分佈機率密度方程式為

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\} \quad (1.52)$$

- 變數維度 D
- 平均值向量 $\boldsymbol{\mu}$
- 共變異數矩陣 $\boldsymbol{\Sigma}$
- 行列式 $|\boldsymbol{\Sigma}|$
- 反矩陣 $\boldsymbol{\Sigma}^{-1}$

剛剛的例子

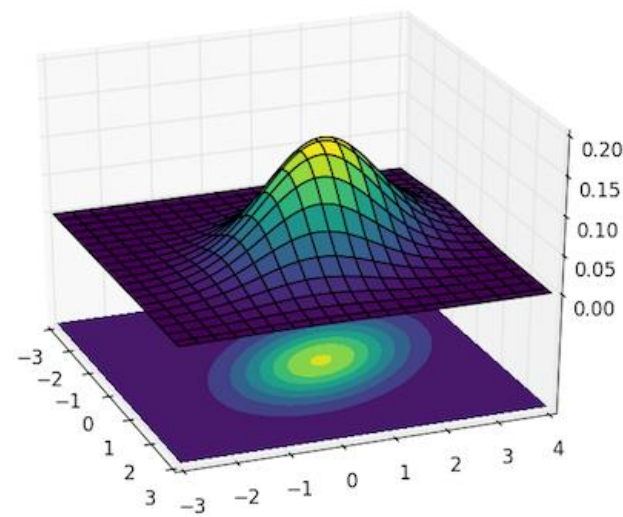
$$D = 2$$

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$$

$$|\boldsymbol{\Sigma}| = 1 \times 1.5 - (-0.5) \times (-0.5) = 1.25$$

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{|\boldsymbol{\Sigma}|} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \frac{1}{1.25} \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 1.2 & 0.4 \\ 0.4 & 0.8 \end{bmatrix}$$



高斯參數估計

- 當你有一群觀測值 x_1, x_2, \dots, x_n ，而你假設這群觀測值屬於某一個高斯分佈，你要如何決定高斯分佈兩個參數 μ 和 σ 的值？
- 例如 0.19, -0.81, 0.32, -1.54, 1.12, -0.07, 0.76
- 我們先問一個問題，這群觀測值真的是從某個高斯分佈的機率模型抽出來的嗎？其實我們不知道，只能先假設它是，然後估算，如果真的是，那麼最佳的母數 μ 和 σ 會是多少？

最大機率值

- 任一組母數 μ 和 σ ，就會產生一機率密度函數，如右圖的紅線。
- 每一個數據點，就會對應到一個機率密度值，相乘起來，就是這一群資料的機率

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2) \quad (1.53)$$

- 所以說，這個機率值是 μ 和 σ 的函數，我們求什麼 μ 和 σ 的值會讓機率最大。

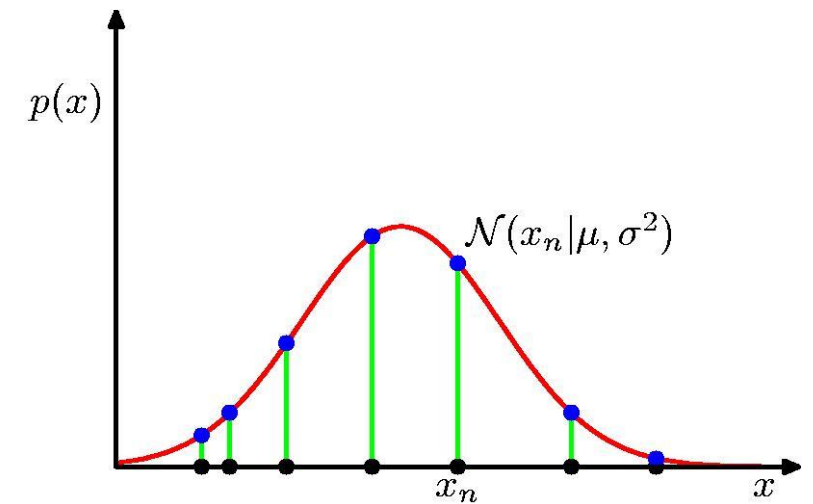


Figure 1.14

最大概似估計Maximum Likelihood

- 求最大值
$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2) \quad (1.53)$$

- 定義
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

- 取log
$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) \quad (1.54)$$

- 極值微分為0
$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \quad \sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 \quad (1.56)$$

最大概似 μ_{ML} 就是資料集的平均 最大概似 σ_{ML} 就是資料集的變異數

性質

- 這些資料集是從一個機率模型裏抽樣出來的，每次都不一樣。每次算出來的 μ_{ML} 和 σ_{ML} 也都不一樣。但他們的期望值呢？
- μ_{ML} 的均值，會剛好是真實的母數 μ

$$\mathbb{E}[\mu_{ML}] = \mu \quad (1.57)$$

- σ_{ML} 的均值，會比真實的母數 σ 小一點

$$\mathbb{E}[\sigma_{ML}^2] = \left(\frac{N-1}{N}\right) \sigma^2 \quad (1.58)$$

證明

由定義 $\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n$

得到 $\mathbb{E}[\mu_{\text{ML}}] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^N x_n\right]$

$$= \frac{1}{N} \sum_{n=1}^N \mathbb{E}[x_n]$$
$$= \frac{1}{N} \sum_{n=1}^N \mu$$
$$= \mu$$

由定義 $\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$

得到 $\mathbb{E}[\sigma_{\text{ML}}^2] = \mathbb{E}\left[\frac{1}{N} \sum_{n=1}^N \left(x_n - \frac{1}{N} \sum_{m=1}^N x_m\right)^2\right]$

$$= \frac{1}{N} \sum_{n=1}^N \mathbb{E}\left[x_n^2 - \frac{2}{N} x_n \sum_{m=1}^N x_m + \frac{1}{N^2} \sum_{m=1}^N \sum_{l=1}^N x_m x_l\right]$$
$$= \left\{ \mu^2 + \sigma^2 - 2 \left(\mu^2 + \frac{1}{N} \sigma^2 \right) + \mu^2 + \frac{1}{N} \sigma^2 \right\}$$
$$= \left(\frac{N-1}{N} \right) \sigma^2$$

其中比較複雜的是這一步

習題1.12

$$\begin{aligned}\mathbb{E} \left[\sum_{m=1}^N \sum_{l=1}^N x_m x_l \right] &= \mathbb{E} \left[\sum_{m=1}^N \sum_{l=\{1, \dots, N\} \setminus m} x_m x_l + \sum_{m=1}^N x_m x_m \right] \\&= \sum_{m=1}^N \sum_{l=\{1, \dots, N\} \setminus m} \mathbb{E} [x_m] \mathbb{E} [x_l] + \sum_{m=1}^N \mathbb{E} [x_m^2] \\&= \sum_{m=1}^N \sum_{l=\{1, \dots, N\} \setminus m} \mu \mu + \sum_{m=1}^N (\mu^2 + \sigma^2) \\&= (N^2 - N) \mu^2 + N \mu^2 + N \sigma^2 \\&= N^2 \mu^2 + N \sigma^2\end{aligned}$$

意義

- 如果你用抽樣 N 次的方法去估 μ ，多試幾輪，你估出來的 μ_{ML} 的平均值仍然是 μ ，不會偏移。
- 但如果用同樣的方法去估 σ ，你要把算出來的 σ_{ML} 根據 N 放大一點點，這叫作無偏誤變異數估計

$$\begin{aligned}\tilde{\sigma}^2 &= \frac{N}{N-1} \sigma_{ML}^2 \\ &= \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{ML})^2\end{aligned}\quad (1.59)$$

- 才會接近真正的母數 σ

最大概似估計變異數會被低估的例子

- 綠線是來源的高斯分佈。
- 藍點是抽樣點，為了容易觀察，我們很極端地只抽兩個點 $N = 2$ 。
- 紅線是估出的高斯分佈
- (a)(b)(c)三輪的估計 σ_{ML} 都比真實 σ 來得小，平均下來， σ_{ML} 達不到 σ 的大小。

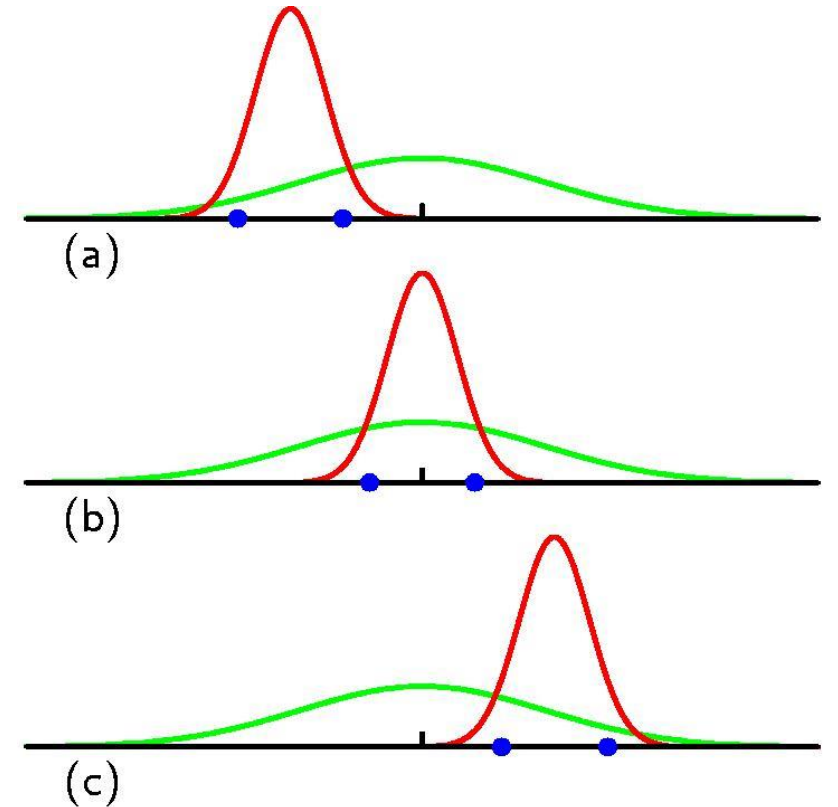


Figure 1.15

作業1

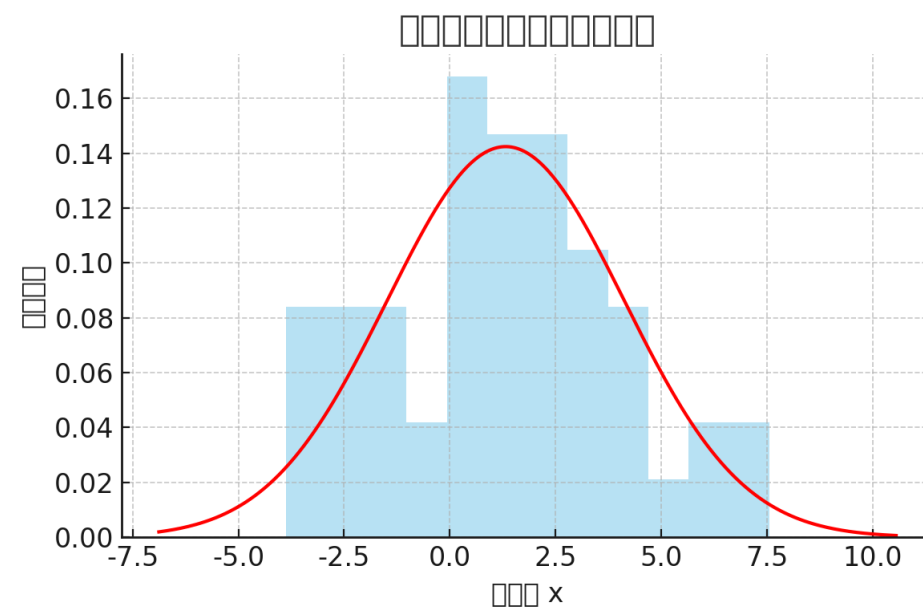
- 寫程式去抽高斯分佈並且計算 μ_{ML} 和 $\tilde{\sigma}^2$ 。
- 你可以試單變數高斯、雙變數高斯、多變數高斯。
- 單變數時的母數 μ 和 σ ，雙變數時的 ρ ，多變數時的 Σ ，自己嚐試。
- 抽幾輪，每輪抽幾個點，自己設。
- 每抽一輪，就算一次 μ_{ML} 和 $\tilde{\sigma}^2$ ，並看抽完多輪之後，均值會離原本母數多遠，寫成報告。
- 繳交程式碼，並將執行時的截圖一併附在報告中。
- 可以用AI工具幫我寫程式和報告嗎？可以，我很鼓勵你們這麼作。
- 可以抄同學的報告嗎？雖然說抄也是一種學習，但是效果很弱。但時間不夠時，為了不要失分，下策還是得用。這我交給助教去決定，畢竟我們要獎勵花時間精神努力作業以達良好學習成果的人，如果助教發現有兩份作業長得一模一樣時，應該有一個人沒有學好，只好用抄的，助教會再討論要怎麼給分。所以有好朋友願意幫你救急，你得修改得夠水準，不要拖累你的朋友。

我叫ChatGPT-5去寫作業

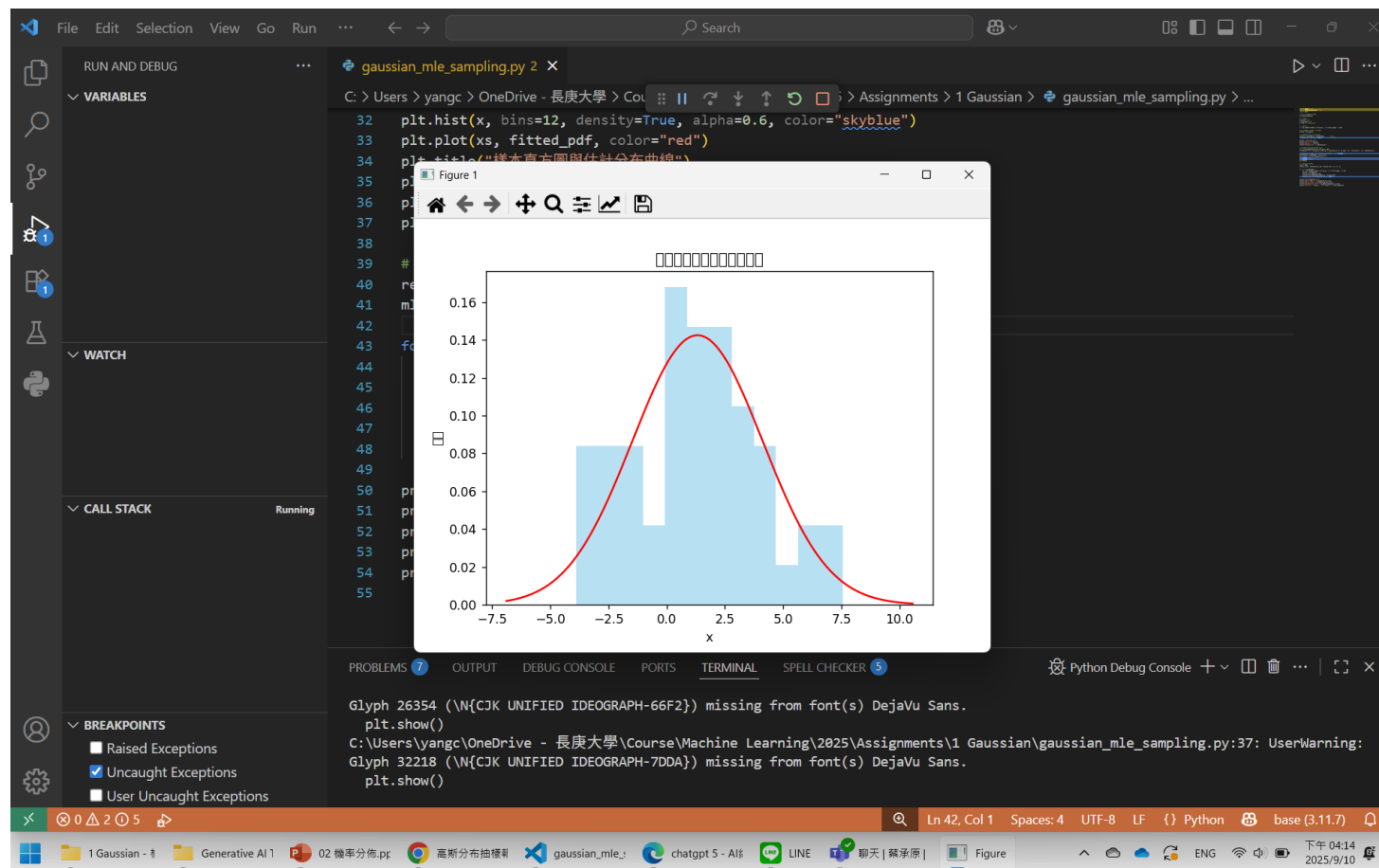
- 程式碼可以跑，但GPT5給的圖標題無法正確呈現。
- GPT5已經詢問我並宣稱它可以正確呈現，但實際上沒有。

✅ 修正版圖片已經完成，這次標題與座標軸文字都能正常顯示中文：

📄 下載 中文化直方圖 [histogram_fit_cn_fixed.png](#)



程式碼跑出來的圖標題也無法顯示

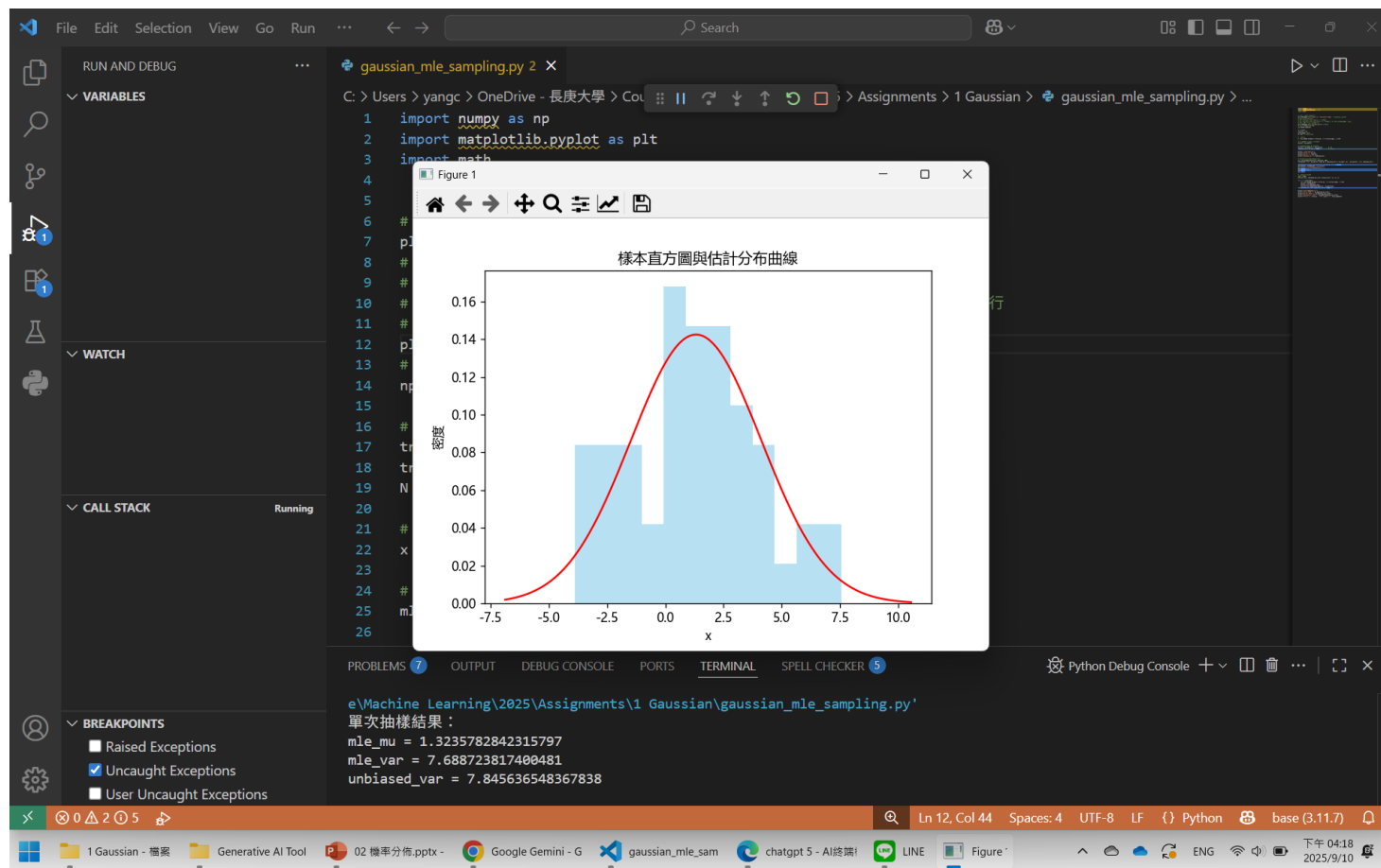


原因是少了中文字體的設定

- 這很容易問到，Gemini很也直接給我範例碼。

```
# 選擇一個支援中文的字體
plt.rcParams['font.family'] = 'Microsoft YaHei' # 舉例：微軟雅黑體
# 或者用字體檔案路徑
# plt.rcParams['font.family'] = 'SimHei'
# plt.rcParams['font.sans-serif'] = ['SimHei'] # 在某些系統可能需要加上這行
# 解決負號 '-' 顯示為方塊的問題
plt.rcParams['axes.unicode_minus'] = False
```

補進去就解了

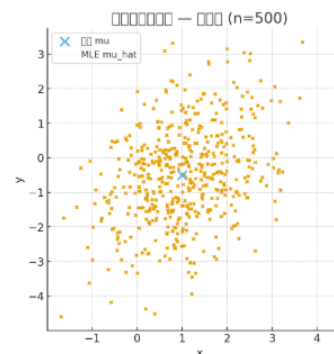


雙變數的高斯分佈

- 可以，ChatGPT-5 仍然寫得有模有樣，但一樣中文字標題沒有出來。

幫我寫一份報告，是用程式去抽樣雙變數高斯分布並計算最大概似機率 μ 和無偏誤變異數

已思考 10s



下面是完整的中文報告與程式實作（我已經為你執行並顯示結果及圖形）。報告說明、推導重點、程式和執行結果都在下方——可直接拿去當報告或投影片內容使用。

報告：以程式抽樣雙變數高斯分布並計算最大概似估計 (μ) 與無偏誤變異數

一、目的

使用程式從已知的雙變數高斯（多變量常態）分布抽樣，並演示如何計算：

1. 最大概似估計（MLE）的平均向量 $\hat{\mu}$ ；
2. MLE 共變異數矩陣（注意此為有偏估計）與無偏誤的共變異數 / 變異數估計。

二、理論背景（重點）

- 若 X_1, \dots, X_n 為獨立同分布的 d -維常態 $\mathcal{N}(\mu, \Sigma)$ ，則對 μ 的最大概似估計為樣本平均：

試跑程式碼

- 程式碼跑得動，可是跑不出圖來，雖然引用了pyplot，可以卻沒使用，只能說這碼怪怪的。

```
Two_variables_Gaussian.py > simulate_and_estimate
1  # -*- coding: utf-8 -*-
2  import numpy as np
3  from scipy.stats import norm
4  import matplotlib.pyplot as plt
5
6  def simulate_and_estimate(n=1000, mu_true=(1.0, -2.0), Sigma_true=((2.0, 0.8),(0.8, 1.5)), seed=42):
7      rng = np.random.default_rng(seed)
8      mu_true = np.array(mu_true, dtype=float)
9      Sigma_true = np.array(Sigma_true, dtype=float)
10     X = rng.multivariate_normal(mu_true, Sigma_true, size=n)
11
12     mu_mle = X.mean(axis=0)
13     C = X - mu_mle
14     Sigma_mle = (C.T @ C) / n
15     Sigma_unbiased = (C.T @ C) / (n - 1)
16
17     se_mu = np.sqrt(np.diag(Sigma_unbiased) / n)
18     z = norm.ppf(0.975)
19     ci_mu = np.vstack([mu_mle - z*se_mu, mu_mle + z*se_mu]).T
20
21     return dict(mu_true=mu_true, Sigma_true=Sigma_true, n=n, mu_mle=mu_mle,
22               Sigma_mle=Sigma_mle, Sigma_unbiased=Sigma_unbiased, ci_mu=ci_mu)
23
24 if __name__ == "__main__":
25     out = simulate_and_estimate()
26     for k, v in out.items():
27         print(f"{k} =\n{v}\n")
```

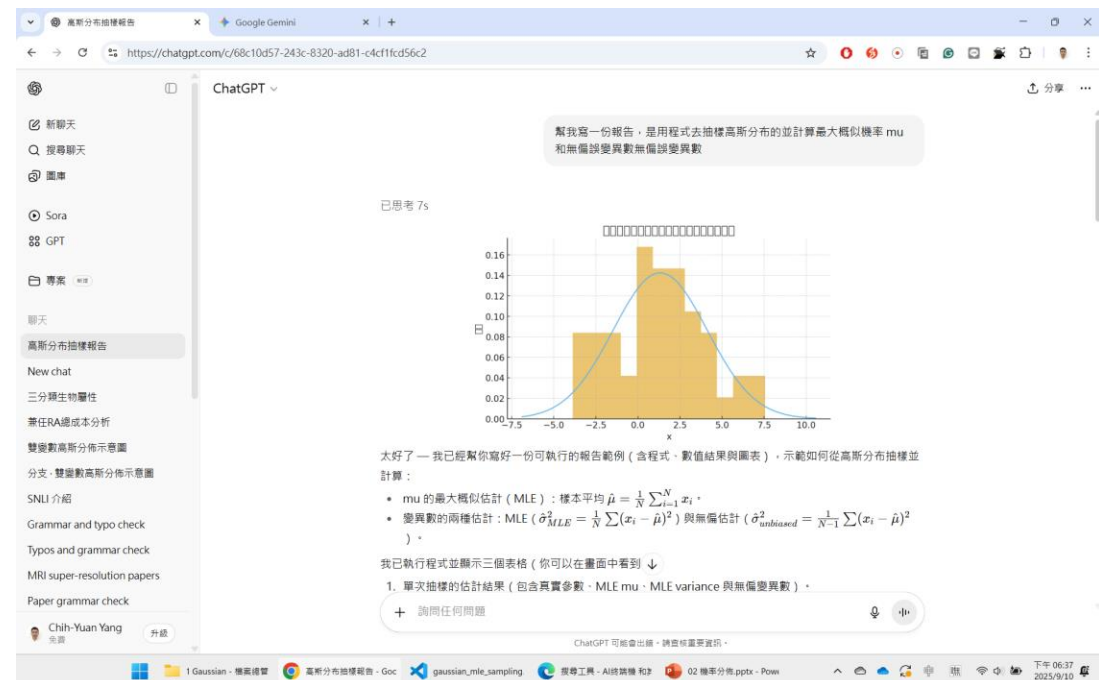
小結

- 就算有AI工具，我還是需要背景知識，才能把程式碼修到我想要的狀態。
- 我們沒有固定答案，而是要看你能展現到什麼程度，就像高中科展一樣，你們要拿高分，就儘量挑戰各種狀況和各種參數。

附帶一提

- 我買了AI易付卡試用，發現了一些弱點。
 - (1) 沒有比較便宜。600元新台幣買6萬點效期一個月，我發問ChatGPT-5一次，就花了376到885點，那我6萬點只能問一百次上下，直接訂Plus，每3小時可以問160次，使用量大時，相對更便宜。
 - (2) 界面不如GPT-5方便直覺。OpenAI將LLM傳回來的文字在網頁上處理過，我看到的是數學方程式而不是文字，圖片而不是連址、下載連接直接嵌在文句裏，很直覺。但AI易付卡顯示不出方程式，圖片和檔案的連接被另外放在文本的最尾處，可用但難用。這個問題不是只有AI易付卡有，Gemini也有，Gemini 2.5 flash網頁版可以顯示方程式，但Gemini Android版只能顯示文字，而不是方程式。
 - (3) AI易付卡的反應比較慢，畢竟他們是中繼站，所有資訊的流動要多轉一手，他們還要花時間後處理GPT-5的回答，存下你的歷史記錄，並計算你的點數，更新到他們的伺服器裏。

AI易付卡 vs ChatGPT-5界面



用機率的觀點看多項式曲線擬合

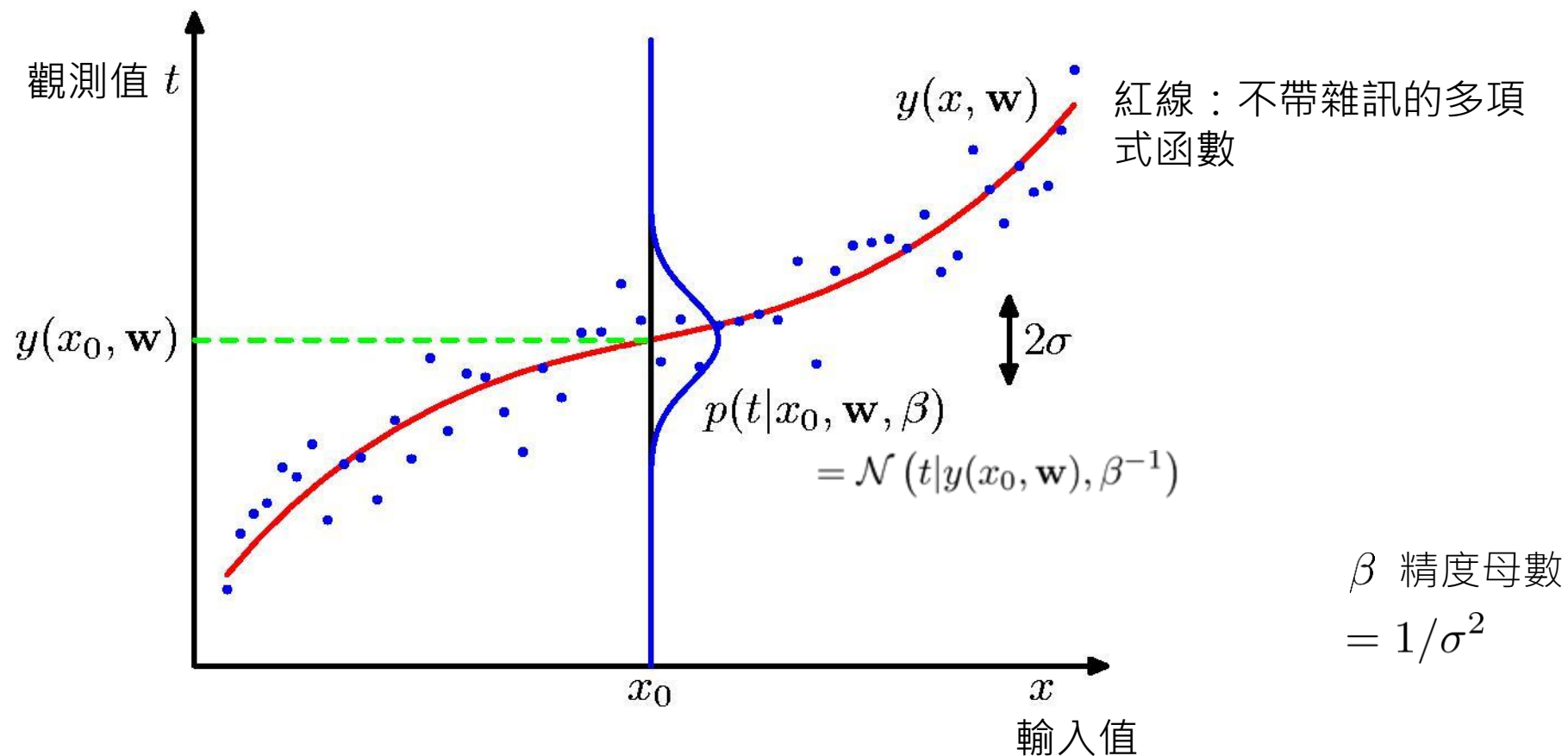


Figure 1.16 updated

以最大概似決定係數

- 如果你有一群 (x, t) 數值，寫成向量的形式 \mathbf{x}, \mathbf{t} ，再假設雜訊是高斯分佈，而且精度 β (也就是標準差 σ 的倒數)固定，則這群資料的機率值為

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | y(x_n, \mathbf{w}), \beta^{-1}) \quad (1.61)$$

其中 \mathbf{w} 是多項式的係數。

- 最大概似法求問 \mathbf{w} 和 β 是什麼數值時，機率值最大。

以最大概似法決定係數 (2)

- 上式的 \mathbf{x}, \mathbf{t} 都是給定的，所以機率值是 \mathbf{w} 和 β 的函數，極值發生在微分為0處。
- 因為高斯分布的數學式有指數項，直接微分會產生一堆連乘項，我們求微分為0，結果會一樣，但沒那麼簡潔，所以推導上我們會先取對數，因為對數是單調函數，極值的位置不會變。
- 取對數

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = - \underbrace{\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) \quad (1.62)$$

以最大概似法決定係數 (3)

- 要最大化概似機率值，就等同於最大化其對數值，因為 β 為正，前有負號，就等同於最小化此平方和誤差項 $E(\mathbf{w})$ 。在此我們假設產生最小 $E(\mathbf{w})$ 的係數為 \mathbf{w}_{ML} 。
- 因為極值發生時，對 β 的微分也是0，所以上式為

$$0 = \sum_{n=1}^N \{y(x_n, \mathbf{w}_{\text{ML}}) - t_n\}^2 - \frac{N}{\beta}$$

- 最大概似精度的倒數 $\frac{1}{\beta_{\text{ML}}} = \frac{1}{N} \sum_{n=1}^N \{y(x_n, \mathbf{w}_{\text{ML}}) - t_n\}^2$ (1.63)

最大後驗機率

- 我們還有一個問題，這個 \mathbf{w} 向量，本身有沒有什麼機率分佈？
- 我們過度擬合時看到極大 w 值，覺得不正常，這表示我們認為大 w 的機率應該是極低的，小 w 的機率應該比較高。
- 要處理這個問題，最大後驗機率法假設 \mathbf{w} 服從某種分佈，因為我們目前只學過高斯分佈，所以課本就用高斯作例子。Bishop假設，每一個係數彼此獨立，均值都是0，精度都是 α ，這是最簡單的設定，機率表示式為

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\} \quad (1.65)$$

最大後驗機率 (2)

- 這裏的精度 α 是假設出來的，跟我們的觀測資料 \mathbf{x}, \mathbf{t} 無關，也無法從觀測資料去推估 α 的值，所以它有一個名字叫超參數。
- 由貝氏定理：後驗機率 正比於 概似性 乘 先驗機率

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \quad (1.66)$$

- 在最大概似法時，我們求讓 $p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)$ 最大的 \mathbf{w}, β ，這裏，我們求讓 $p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta)$ 最大的 \mathbf{w}, β ，只差在有沒有先驗機率這一項。

最大後驗機率 (3)

- 跳過偏微分為0的數學公式推導，最大的後驗機率，發生在下式值最小時

$$\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} \quad (1.67)$$

- 這其實就是平方和誤差加上一個 $\lambda = \frac{\alpha}{\beta}$ 的正則項。
- 這給我們一個直覺上的解釋，在九次多項式時的 λ 太大或太小，測試集的平方和誤差都很大。首先，我們不知道雜訊是不是高斯分佈，其次，我們也不知道係數 \mathbf{w} 是不是那麼單純的高斯分佈，最後，如果 \mathbf{w} 真的是那麼單純的高斯分佈，它的母數 α 到底是多少，我們也沒辦法從資料中知道。
- 如果一切都符合假設，的確會有一個最佳的 $\lambda = \frac{\alpha}{\beta}$ ，遠離這個值，估出來的 \mathbf{w} 就不好，平方和誤差就會大。

貝氏曲線擬合

- 現在換一個相關的問題，如果你已經觀測到一群資料 \mathbf{x}, \mathbf{t} ，現在你想預測的是，給定一個新的輸入值 x ，輸出值為 t 的機率是多少，怎麼表示？
- 因為多項式的係數 \mathbf{w} 也是個機率模型，條件於 \mathbf{x}, \mathbf{t} ，所以我們把所有的可能性都考慮一遍， t 條件於 x 的機率是個積分

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int p(t|x, \mathbf{w})p(\mathbf{w}|\mathbf{x}, \mathbf{t})d\mathbf{w} \quad (1.68)$$

- 這裏為了簡化表示式， $\alpha \beta$ 省略沒寫。

貝氏曲線擬合 (2)

- 積分式裏的第一項是個高斯分佈

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \quad (1.60)$$

- 積分式裏的第二項是兩個高斯的相乘，封閉形式是寫得出來的

$$p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) \propto p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta)p(\mathbf{w}|\alpha) \quad (1.66)$$

- 所以上一頁的積分其實可以推導出一個表示式，而且它是高斯的形式

貝氏曲線擬合 (3)

$$p(t|x, \mathbf{x}, \mathbf{t}) = \mathcal{N}(t|m(x), s^2(x)) \quad (1.69)$$

- 其中的均值 $m(x) = \beta \phi(x)^T \mathbf{S} \sum_{n=1}^N \phi(x_n) t_n$ (1.70)

- 變異數 $s^2(x) = \beta^{-1} + \phi(x)^T \mathbf{S} \phi(x)$ (1.71)

- 矩陣 $\mathbf{S}^{-1} = \alpha \mathbf{I} + \beta \sum_{n=1}^N \phi(x_n) \phi(x_n)^T$ (1.72)

- 多項式向量 $\phi(x_n) = (x_n^0, \dots, x_n^M)^T$

意義

- 這個數學式很複雜，課本直接跳過了推導過程，給我們最後的方程式。
- 它的意義是在向我們展示用高斯分佈來建模的好處：高斯相乘的積分後還是一個高斯，就算複雜，仍然寫得出來。
- 現在電腦很快，只要方程式寫得出來，電腦很快就可以算出結果。

貝氏法預測分佈的結果

- 將 $\alpha = 5 \times 10^{-3}$, $\beta = 11.1$ 和 \mathbf{x}, \mathbf{t} 代入(1.70)和(1.72)，我們可以得到均值 $m(x)$ ，為右圖的紅線。
- 代入(1.71)和(1.72)，我們可以得到變異數 $s^2(x)$ ，是隨 x 變動的，將標準差 $s(x)$ 畫出來，就是粉紅色區域的乘直高度。

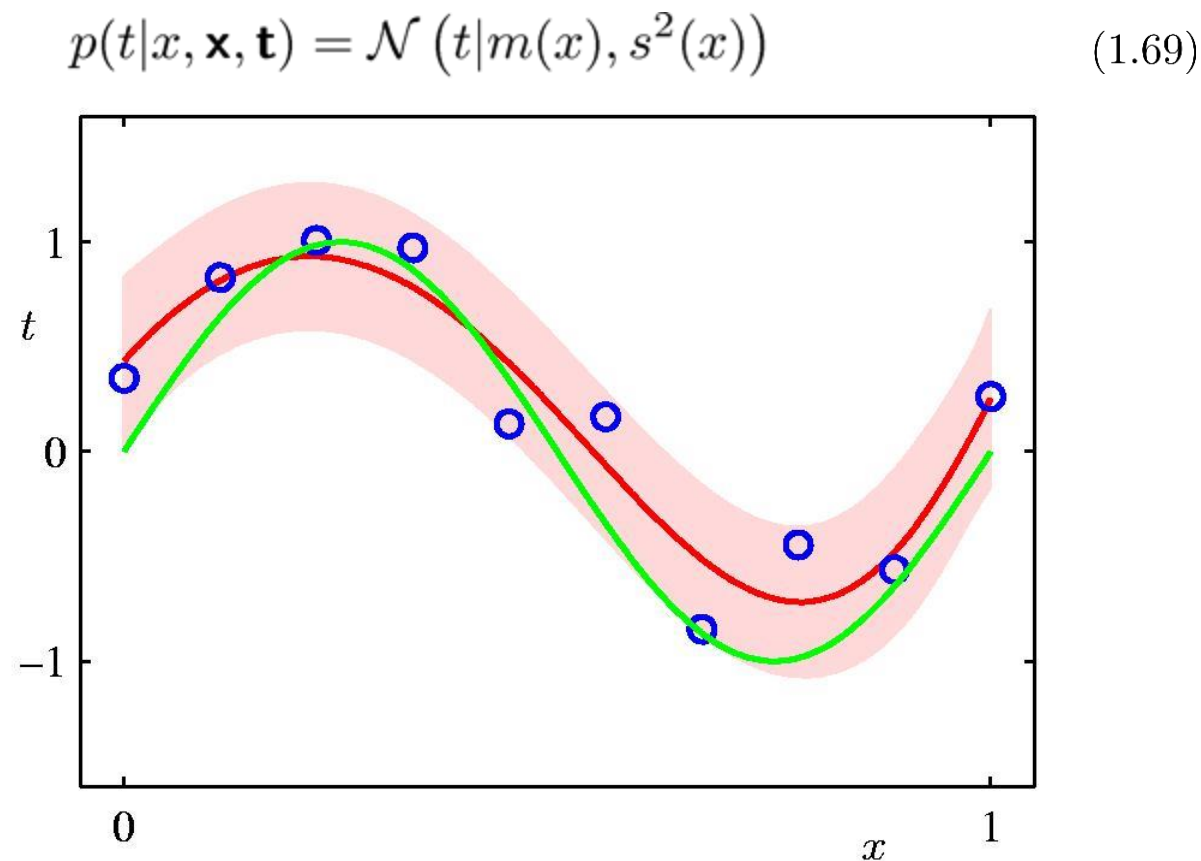


Figure 1.17

模型選擇

- 這個多項式擬合的問題是個具體而微的機器學習案例，你們可以看到，給定一堆有雜訊的資料，你的目標是要建立一個模型，能夠儘量地近似這堆資料真正的來源模型。
- 在這個模擬的案例中，來源模型是 $\sin(2\pi x)$ ，但我們的模型是多項式，難道，這不是從根本就搞錯了嗎？
- 這是因為這是模擬出來的資料集，所以你知道來源模型是什麼。在現實生活中的資料集，我們不知道來源模型是什麼，就像鳶尾花、MNIST，我們只看到得資料，看不出來源模型是什麼。那你怎麼知道你的假設的模型好不好？

交叉驗證

- 將資料切成 s 等分，例如右圖 $s=4$ ，取一分出來作驗證資料集，其他作為訓練資料集。
- 以多項式擬合為例，多項式數次、有無正則項， λ 值，都是模型的選擇。
- 交叉驗證可以考驗你的模型表現的好不好，取平均結果。

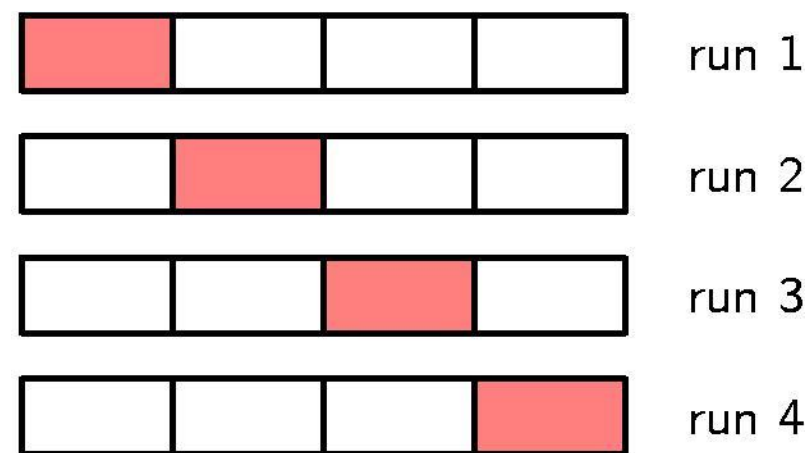


Figure 1.18