
機器學習與數據分析
商汤
sensetime

鑽石的價格

鑽石恒久遠，一顆永流傳。

Q：鑽石多少錢一克拉？

鑽石價格的影響因素： 重量、色澤、淨度、切割的形態 ……

行家看一下鑽石的大小和色澤，就能很快判斷它的價格。行家是怎麼做到的呢？

鑽石的價格

鑽石專家

無數鑽石

成交的價格

經驗的積累和總結

- 總結出鑽石的定價規律
- 價格如何受各種因素影響

這是一個從大量觀察中學習的過程。

在過去學習是高等動物特有的能力。

隨著人工智能技術的發展，可以讓機器也具備學習的能力。

機器學習，就是讓機器模擬人的學習行為，從觀察或者實踐中獲得判斷、預測、或推理能力的過程。

預測

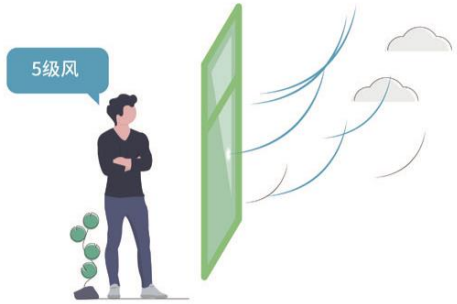
鑽石專家根據鑽石的重量和其它屬性來推測它的價格。

根據觀察到的屬性去推測相關數值的過程，我們稱之為預測（predict），也叫推斷(inference)。預測是智慧的一種重要表現形式。

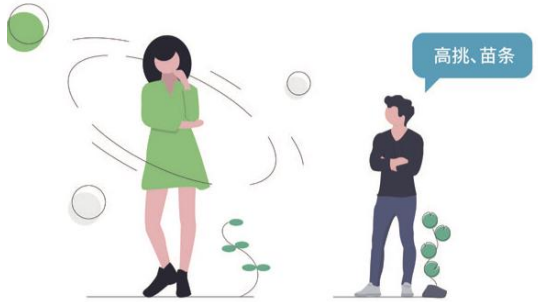
Q：我們生活中有哪些預測？

圖：判斷別人是否開心

圖：判斷菜好不好吃



圖：通過聲音判斷風力



圖：通過體型、身高判斷體重

Q：在這些預測裡，我們使用了什麼樣的屬性？又是如何根據這些屬性推測結論的呢？

數字化：建立機器與現實世界的連接

人 能通過表情、菜肴、 風聲等等來進行預測。

機器：？？？

機器的世界裡面只有數字。

讓機器進行預測

數字化 建立 現實世界 / 機器世界 的連接， 用數字來表達觀察到的屬性。

通過數字化的過程，機器就可以對需要進行預測的事物建立數字表達。

不同的屬性——不同的數字化方式

一顆鑽石有幾個不同方面的屬性，包括**重量**、**淨度**、和**切割形態**等。這些不同的屬性數字化方式是不一樣的。

- **重量**的數值是可以直接通過天平稱量獲得的(鑽石通常以克拉為單位)；
- **淨度**則根據內含物的量，可以分成從無暇級到內含級等若干個等級。

在數字化的過程中， 可給不同的等級賦予從小到大的不同整數。

數字化：建立機器與現實世界的連接

表 鑽石淨度等級的數字化表示

等 級	描 述	數字表示
無暇級(FL)	鑽石沒有任何內含物和表面特徵	0
內無暇級(IF)	鑽石無可見內含物	1
極輕微內含級 (VVS)	鑽石內部有難以發現的極微小內含物	2
輕微內含級(VS)	鑽石內部可以看到微小的內含物	3
微內含級(SI)	鑽石有可見的內含物	4
內含級(I)	鑽石的內含物明顯可見，影響透明與光澤	5

經過數字化之後，一個事物就可以由幾個數字表達。

一顆 0.5 克拉，淨度為輕微內含級的鑽石就可以表示為 (0.5,3) 。

通常我們稱這樣一組按照特定順序排列的數字為**向量**，而其中數字的個數稱為向量的維數。

向量 (0.5,3)，它的維數是2，我們稱之為**二維向量**。

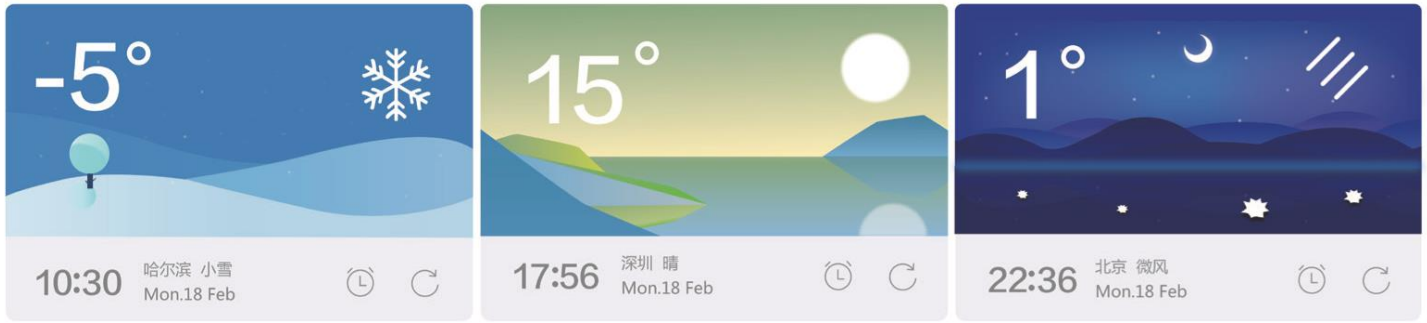
數字化：建立機器與現實世界的連接

1.請依據上述鑽石淨度等級的數字化方案，用向量表示下面的鑽石:

a) 1 克拉的無瑕級鑽石

b) 0.3 克拉的微内含級鑽石

2. 每天的天氣是我們在生活中每天都需要關心的。



一般來說，天氣報告會給我們提供關於天氣的幾個主要指標:

a)天氣狀態(晴、少雲、多雲、陰天、下雨)；b)氣溫；c)風級；d)相對濕度

請根據上述指標，用一個四維向量表示當前的天氣。

預測函數：從觀察到結論的映射

對事物進行了數字化，然後機器就可以對事物進行預測。

人類在做判斷和預測時，要經過一個複雜的心理過程，很難在只懂得數字的機器中直接顯示。

現代人工智能**把預測簡化為一個數學函數**，使之可以在機器上通過演算法實現。這個函數通常被稱為**預測函數**。

函數

函數是一種對應法則，對於一定範圍內的每一組輸入數值，通過函數都會對應一個唯一的輸出值。很多函數都可以通過計算公式或者曲線圖表達。

在初中數學裡面，我們學習過很多初等函數：正比例函數、反比例函數、一次函數與二次函數。

預測函數：從觀察到結論的映射

應該用什麼樣的函數做預測呢？

預測函數通常是因問題而異的。一般通過看預測結果是否符合實際情況來判斷預測函數是否合適。

預測函數的結果
- 符合實際情況 √ 合適
- 不符合實際情況 × 不合適

也就是說，**一個好的預測函數，它的預測結果應該是和實際情況相吻合的。**

預測的例子 —— 鑽石的重量和價格

重量	價格
3	30
3	20
1	7
3	18
1	5
3	25
1	1
1	4
2	16
2	20
3	21
2	8
3	24
2	6
2	10
4	50

在特定的淨度和切割形態的條件下，鑽石的價格主要是取決於它的重量(克拉數)的。左邊的表格展示了16個不同的鑽石的重量與價格。

Q：能從表中發現重量與價格有什麼規律嗎?

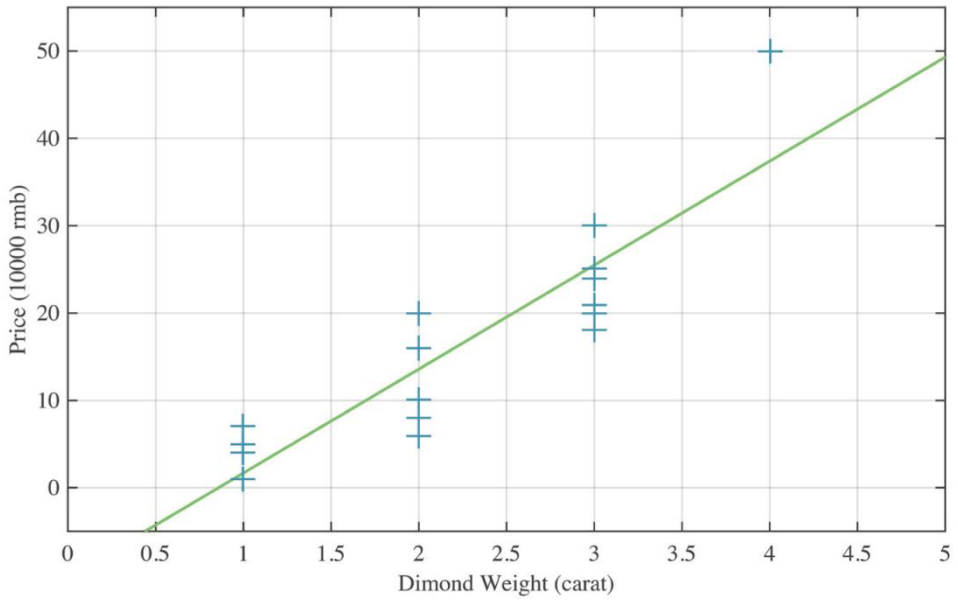
在特定的淨度和切割形態的條件下，

- 鑽石的重量越大，價格越高。

但並不容易歸納出一條數學公式來表達兩者直接的準確關係。

借助**圖表**以更加直觀地展示數值之間的關係，從而幫助我們更有效地建立預測函數。

預測的例子 —— 鑽石的重量和價格



散點圖

- 以橫軸表示鑽石的重量，
- 以縱軸表示鑽石的價格，
- 每個“+”號代表一顆鑽石的座標點， 座標為(重量，價格)。

綠線是透過這些鑽石資料建立的預測函數，用來表示鑽石重量與價格間的關係。

Q：這個預測函數有什麼特點？它像是我們學過的哪種函數？

在這個散點圖中，價格和重量呈現出了一種條帶狀的關係。

因此，我們可以把預測函數設定為**一次函數**，也就是：

$$y=ax+b$$

這裡，x 和 y 分別代表輸入(重量)和輸出(價格)。這個函數裡面還涉及兩個量：斜率 a 和 截距 b，這些在定義預測函數時需要確定的數值稱為參數。

SenseStudy實驗

SenseStudy課程平臺“人工智能入門（上）”

實驗5 – 2 多項式回歸模型

```
train_x = [27,29,34,40,42,47,48,49,50,52,52,52,54]
train_y = [6,7.5,9,10.7,12.8,15.1,16,18.5,19.4,18.4,19.7,21.8,21.7]
fig() + scatter(train_x, train_y)

model = linear_regressor()
model.train(train_x, train_y)
model.show()

x=40
pred_y = model.predict(x)
print(pred_y)
```

SenseStudy實驗

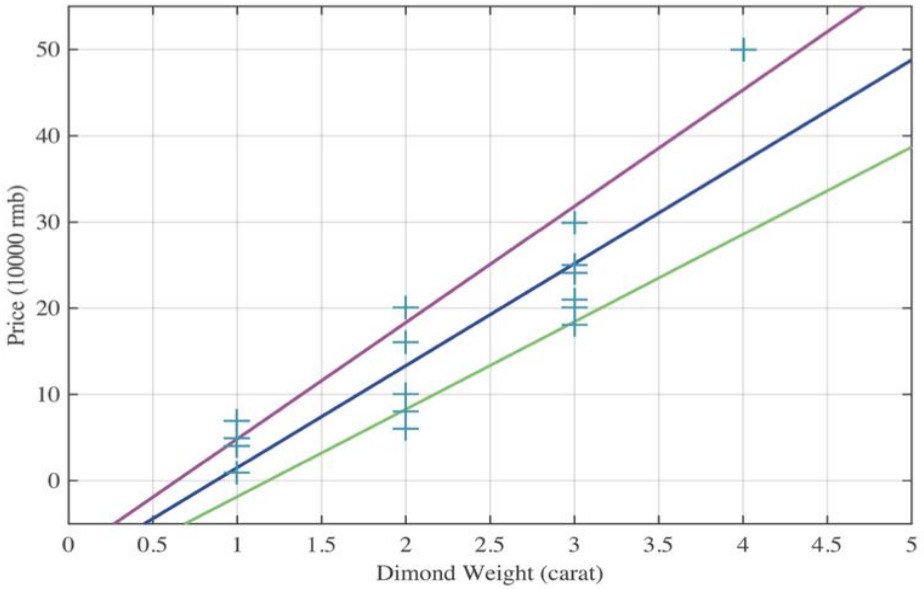
SenseStudy課程平臺“人工智能入門（上）”： 實驗5 – 2 多項式回歸模型

```
train_x = [27,29,34,40,42,47,48,49,50,52,52,52,54]
train_y = [6,7.5,9,10.7,12.8,15.1,16,18.5,19.4,18.4,19.7,21.8,21.7]

model = linear_regressor()
model.train(train_x, train_y)
model.show()
x=40; pred_y = model.predict(x); print(pred_y)

model = poly_regressor(2)
model.train(train_x, train_y)
model.show()
x=40; pred_y = model.predict(x); print(pred_y)

model = poly_regressor(30)
model.train(train_x, train_y)
model.show()
x=40; pred_y = model.predict(x); print(pred_y)
```

對預測函數的評價

對於同一個問題，可以嘗試很多不一樣的預測函數。

當我們把斜率 a 和截距 b 設為不同的值，就可以得到不同的預測函數(如右圖所示)。



Q：你認為圖中的哪個預測函數較為準確？你的依據是什麼？

通過**直觀觀察**發現，藍色直線表示的預測函數和實際資料是基本吻合的。

在實際應用中，我們需要一種**數值化的方式**來評價一個預測函數的準確度——這就像老師通過測驗或者考試的分數來評價我們在一個階段的學習成果。

對預測函數的評價

1. 給預測函數輸入一個重量值，讓它預測相應的價格；
2. 將預測值與實際價格(標準答案)相比較，看看有多大的偏差。

這裡，預測值和實際值之間的差別稱為預測誤差，可以通過下面的數學公式計算：

$$\text{預測誤差} = (\text{預測值} - \text{實際值})^2$$

在人工智能領域的實踐中，我們通常會準備**一組樣本資料**對預測函數進行測試，這組用於測試的樣本資料稱為**測試集**。

平均預測誤差:預測函數分別對測試集的每一個樣本進行預測，最終計算出在所有樣本上預測誤差的平均值，用於整體評價這個預測函數。

平均預測誤差越小代表預測越準確。

SenseStudy實驗

SenseStudy課程平臺“人工智能入門（上）”：實驗5 – 3 線性回歸模型評估與測試集

```
train_x = [27,29,34,40,42,47,48,49,50,52,52,52,54]
train_y = [6,7.5,9,10.7,12.8,15.1,16,18.5,19.4,18.4,19.7,21.8,21.7]
model = linear_regressor()
model.train(train_x, train_y)

def compute_error(model, x, y):
    pred = model.predict(x)
    error = 0
    for i in range(len(pred)):
        error = error + (y[i]-pred[i])**2
    error = error / len(pred)
    return error

print(compute_error(model, train_x, train_y))
model2 = poly_regressor(3)
model2.train(train_x, train_y)
print(compute_error(model2, train_x, train_y))
model3 = poly_regressor(30)
model3.train(train_x, train_y)
print(compute_error(model3, train_x, train_y))
```

SenseStudy實驗

SenseStudy課程平臺“人工智能入門（上）”： 實驗5 – 3 線性回歸模型評估與測試集

```
test_x = [23,31,32,38,40,45,49,50,50,51,51,53,55]

test_y = [6.3,7.2,9.1,10.5,12.9,15.5,15.9,18.3,19.7,18.9,19.3,21.3,22.1]

print(compute_error(model, test_x, test_y))

print(compute_error(model2, test_x, test_y))

print(compute_error(model3, test_x, test_y))
```

機器進行預測的過程

(1) 數字化：把一個事物的屬性通過數字表示出來，從而讓機器可以處理；

(2) 預測函數：根據輸入的屬性，計算出預測的結論。

利用評價方式衡量一個預測函數的好壞，如計算平均預測誤差。

確定預測函數參數值的方法

在鑽石價格預測的問題中，我們把預測函式定義成形式為 y=ax+b 的 一次函數，並且通過簡單的幾何計算確定了它的參數：斜率 a 和截距 b。

Q：通過畫圖和幾何計算的方法確定參數值的方式在實際應用中會有什麼問題呢？

通過畫圖計算這種簡單方法來確定參數在實際中是存在很多困難的。

可以自動確定參數的方法 —— **監督學習**。

監督學習是機器學習中一類非常重要的方法，並且在人工智能的實際應用中得到廣泛運用。

監督學習概覽

機器學習的基本想法是從給定的資料中，通過一定的演算法，自動尋求最優的參數設定。這樣的過程我們也通常稱為**訓練**，它包含三個重要方面:

1.訓練數據

- 資料是訓練的基礎。
- 在一般的監督學習中，訓練資料一般包括**多組含有輸入和輸出值的監督樣本**。
- 圖像等訓練資料還需包含類別資訊。

2.訓練目標

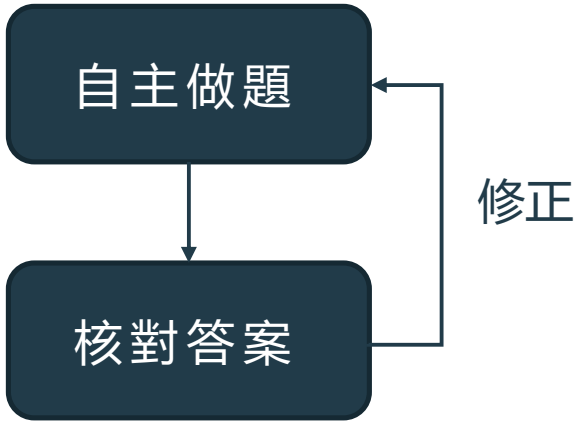
- 機器學習需要一個目標的指引。
- 監督學習的目標可以設定為讓訓練樣本上的平均誤差降到最低。

3.訓練演算法

- 有了目標之後，我們還需要有具體的步驟朝著目標前進。
- 在訓練集上，朝著一定的目標，尋求最好的參數設定一系列步驟， 就構成了訓練演算法。

試驗與修正：一種簡單的學習策略

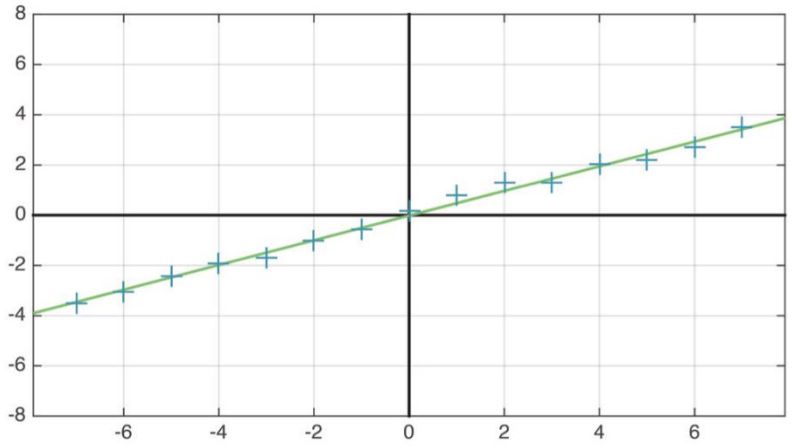
- 學習是一個不斷提高自己的過程。



我們在學習過程中經常運用一種有效的策略——**試驗與修正**。

這也是機器學習中常見的策略。

“試驗與修正” 策略在電腦上的實現案例



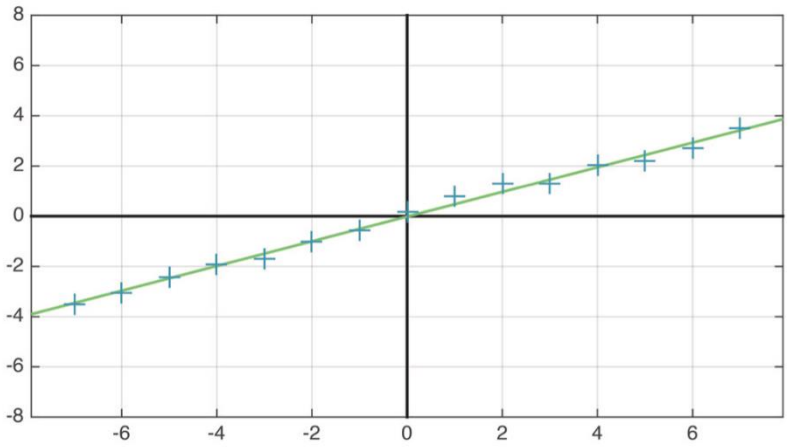
- 符號“+”代表資料的分佈
- 綠線代表根據資料得到的預測函數

Q：仔細觀察，輸出 y 和輸入 x呈現了怎樣的關係?

- 輸出 y 和輸入 x 呈現過原點的線性關係。

- 我們可以設定預測函數的形式為 y=ax。
- 這個函數有一個參數 a，可以通過監督學習尋求一個最優的 a 值。

“試驗與修正” 策略在電腦上的實現案例



- 符號“+”代表資料的分佈
- 綠線代表根據資料得到的預測函數

預測函數的形式為 y=ax

通過監督學習尋求一個最優的 a 值：基於試驗與修正策略的演算法

給定一個訓練集：(x_1, y_1)，…，(x_n, y_n)。每次選取一個點 (x_i, y_i)，進行以下的操作:

1. （試驗）計算預測誤差 $e_i = y_i - ax_i$
2. （修正）對參數 a 進行修正 $a \leftarrow a + \lambda e_i x_i$

反覆運算演算法

(其中 λ 是一個很小的係數，用於控制更新的大小，通常被稱為步長係數。)

上述的操作不斷重複進行，直至到達重複次數的上限或者平均誤差降到預定的水準。

經驗

歸納

新的資料

輸入

規律

預測

未來

歷史資料

訓練

新的資料

輸入

模型

預測

未知屬性

資料對性能的影響 —— 訓練集的大小

從數據中學習 → 機器學習是建立在資料的基礎上的

Q：資料是如何對智慧產生影響的呢？

小實驗：　隨機選取 diamonds 資料集的不同大小的子集訓練預測函數(每種大小的子集隨機選取 3 個來做實驗)，並把預測函數畫出來。

實驗結果：(1)預測函數會擬合，或者說自動適應到訓練樣本上。訓練集越小,預測函數隨機擾動的幅度就越大。

(2)隨著訓練集逐漸增大，得到的預測函數會靠攏到一個固定的位置。

實驗推論：　要獲得一個穩定可靠的預測函數，我們需要一個足夠大的訓練資料集。

Q：訓練資料集太小會讓預測函數不穩定。但訓練集是越大越好嗎？

資料對性能的影響 —— 訓練集的分佈

除了適當的訓練集大小，訓練樣本的分佈對於預測函數也有重大的影響。

Q：資料集的分佈是如何對智慧產生影響的呢？

小練習：diamonds 和 diamonds.b 是兩組不同的鑽石的重量與價格資料，它們是從不同切割形態的鑽石上收集的資訊。之前在 diamonds 上已經訓練好了一個預測函數，請對照 diamonds.b 資料集畫出這個預測函數。它在 diamonds.b 上是一個合理的預測函數嗎?

實驗結果：不合理。在一個資料集上訓練的預測函數，用於其它條件下採集的資料，其預測可能就會發生很大的偏差。

實驗推論：在訓練時選取一個和實際應用分佈類似的資料集是非常重要的。

建立人工智能模型的三個階段

在實際工作中，為了保證一個人工智能模型(比如預測函數)的準確性和可靠性。一般需要經過三個階段：

1.訓練

- 收集一個大小分佈合適的訓練資料，並在其基礎上訓練模型。

2.測試

- 在另一組資料上，測試訓練出來的模型的效能（比如平均預測誤差）。
- 為了保證測試評價的客觀性，測試樣本應該和訓練樣本是分離（沒有重複）的，而且測試樣本的分佈應該反映實際應用時的分佈。

3.部署

- 測試通過(比如性能指標達到預定的標準)的模型，就可以部署在實用環境中使用了。

回歸與分類

Q：下列預測的目標有什麼區別？

鑽石的價格、一個人的體重 → 連續的數值 回歸

某人是開心還是生氣、一盤菜肴是否美味 → 不連續 分類

利用**輸出類型**來判斷

特性	回歸	分類
輸出類型	連續資料	離散數據
目的	找到最優擬合	尋找決策邊界

預測明天的氣溫是多少度，是一個 回歸 任務；

預測明天是陰、晴還是雨，是一個 分類 任務。

預測鑽石的淨度等級 ，是一個 分類 任務。

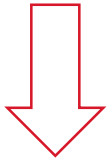
分類任務案例——鑽石的淨度等級

表：不同鑽石的價格、重量和淨度等級

價格	重量	淨度
30	3	高
20	3	高
18	3	低
21	3	低
25	3	高
24	3	高
16	2	高
20	2	高
7	1	高
8	2	低
14	2	低
5	1	高
3	1	低
30	4	低
45	4	高

Q：觀察左側表格，說說淨度的等級 與 鑽石的價格和重量的關係是怎樣的？

鑽石的價格越高，重量越低，對應的淨度等級越高。



可以粗略通過鑽石價格和重量的比值，

來判斷一個鑽石的淨度等級。

分類任務案例——鑽石的淨度等級

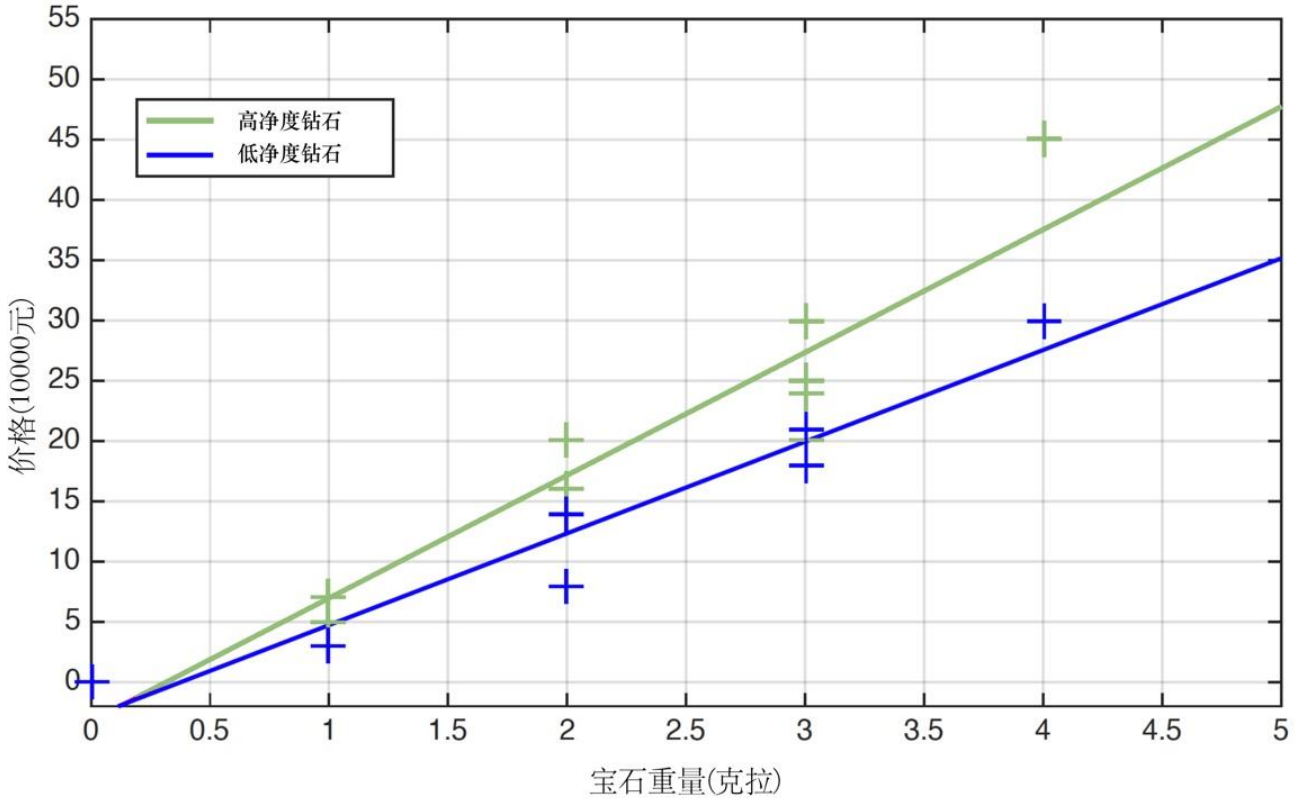


圖　鑽石的價格、重量和淨度等級的散點圖

用散點圖來更形象地表示這些資料

通過將高低淨度的鑽石的座標點畫成不同的顏色。

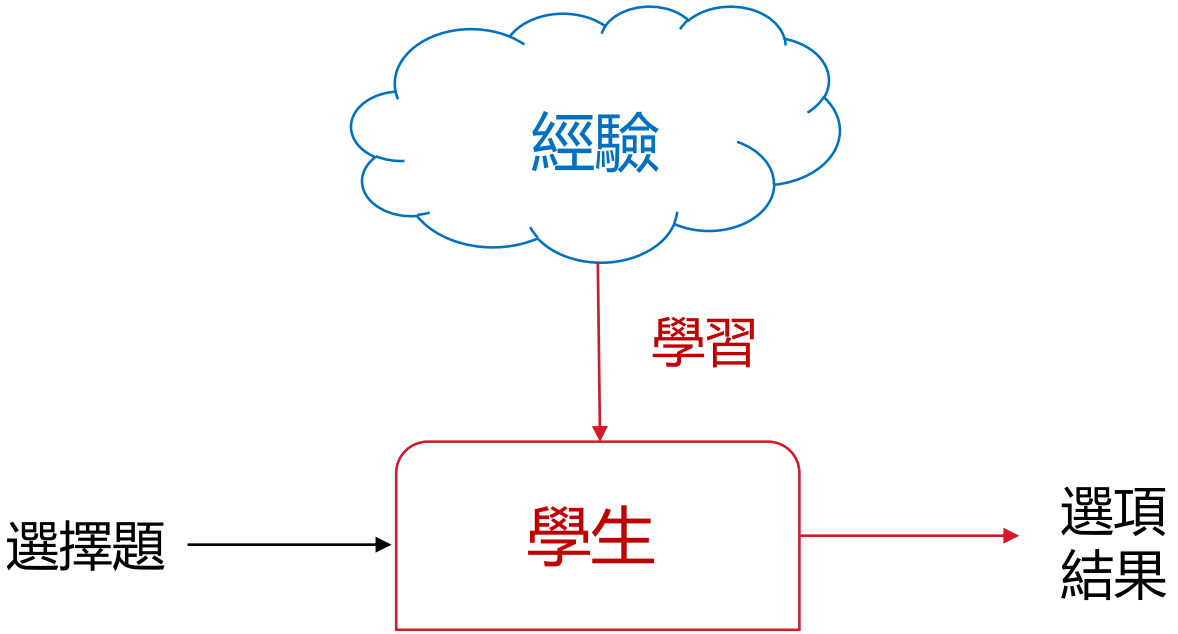
可以發現兩類鑽石明顯處於圖像中兩條不同的直線上 。

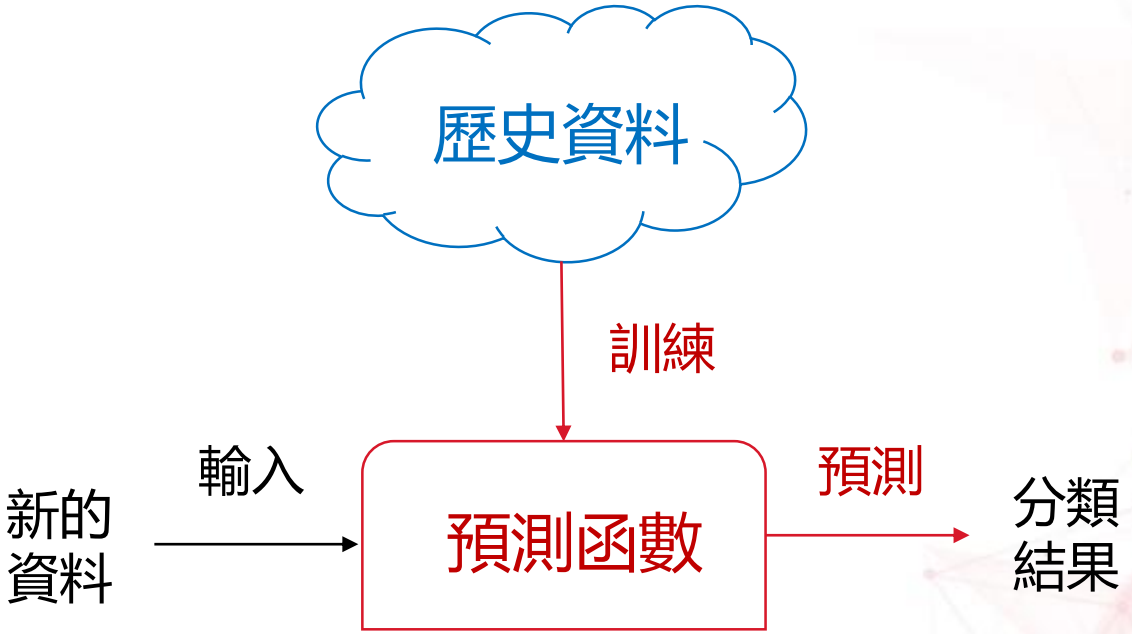
- 綠線代表高淨度鑽石的預測函數；
- 藍線代表低淨度鑽石的預測函數。

綠線斜率大，說明相同重量下，高淨度鑽石價格比低淨度鑽石要高。

我們可以通過觀察和計算鑽石更接近哪條直線，來判斷一個鑽石的淨度係數。

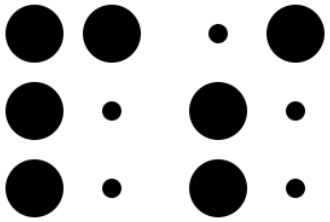
對分類預測函數的評價





教師判別分數

不同學生得分不同



評價 分類預測函數

不同函數準確度不同

對分類預測函數的評價

預測的類別和實際類別的差異被稱為預測誤差，有時候也被稱為分類誤差。

$$\text{預測誤差} = \begin{cases} 0, \text{預測類別} = \text{實際類別} \\ 1, \text{預測類別} \neq \text{實際類別} \end{cases}$$

評價步驟

1.收集一組測試集;

2.在一批樣本上，對這個分類器的誤差進行測試;

3.計算誤差的平均值，也被稱作預測錯誤率。

用於整體評價一個分類預測函數的分類性能。

錯誤率越小的分類器代表預測越準確。

Q：反過來，如果想求得“預測準確率”，請寫出相應的運算式。

SenseStudy課程平臺“人工智能入門（上）”

實驗5 – 5 線性分類模型

```
train_x = [60,56,60,55,60,57,65,60,62,59,43,52,41,45,43,50,46,52,56,56]
train_y = [1,1,1,1,1,1,1,1,1,1,-1,-1,-1,-1,-1,-1,-1,-1,-1,-1]
model = linear_classifier()
model.train(train_x,train_y)
model.show()

x=60.5
pred_y = model.predict(x)
print(pred_y)
```

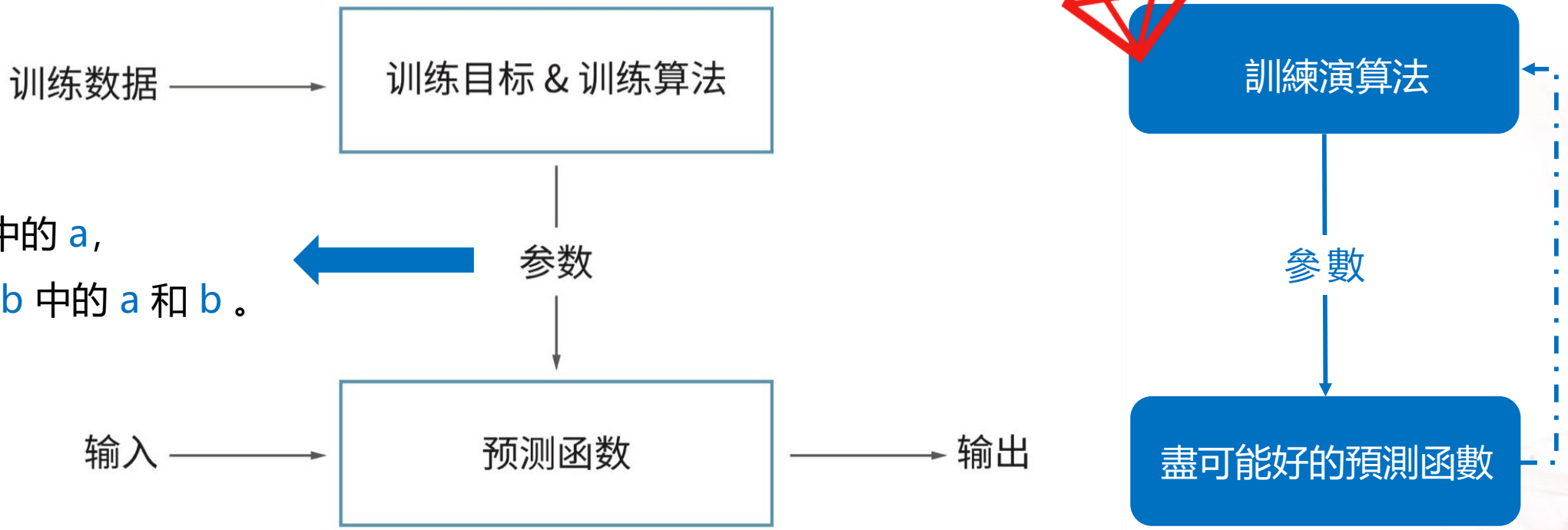
對分類預測函數的評價

Q：如果多種分類規則把四顆鑽石都判斷對了，這意味著大家的分類器都一樣好嗎？

可以類比學生的考試，兩個同學得到了一樣的分數，不代表他們對於知識的掌握程度完全一樣。只能說，在這份試題上，他們的表現一樣好。

所以，這不意味著大家的分類器都一樣好。
只能說，在這個四顆鑽石的測試集上，大家的分類器表現得一樣好。

分類器的封裝和調用



比如 y = ax 中的 a，

比如 y = ax+b 中的 a 和 b 。

相似的問題可以使用同樣的機器學習演算法。

如果將這些演算法進行提前打包，就可以直接使用相應的工具包，從而節省大量的開發時間。

為了方便電腦的開發者，已經定型的演算法有一些成熟的套裝軟體。

綫性分類器的局限性

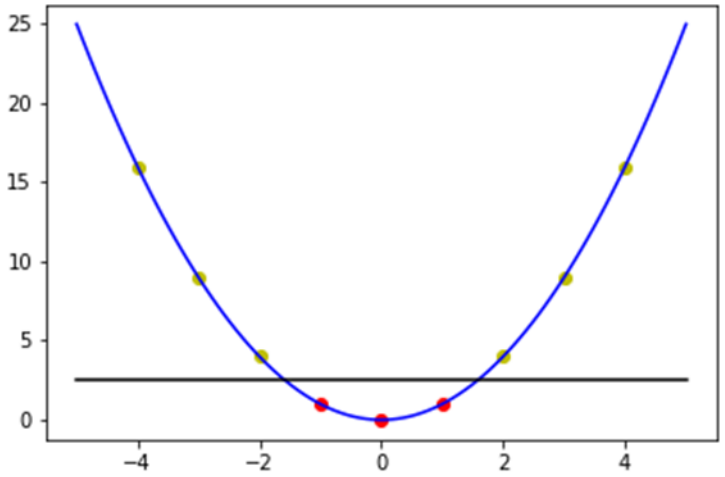
Q：對於一些二分類問題，使用線性分類器無法分類（一條直線無法分開）。比如：第一類：-1，0，1　第二類：-4，-3，-2，2，3，4

需要首先將線性不可分數據轉換成為線性可分數據。

方法：空間變換，一維空間→二維空間，

即 x → (x, x^2)

變換後數據為：第一類：(-1,1) (0,0) (1,1)　第二類：(-4,16) (-3,9) (-2,4) (2,4) (3,9) (4,16)



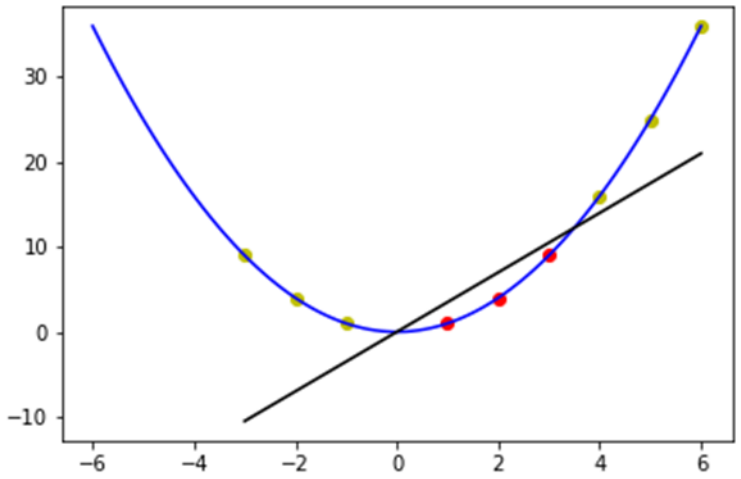
如圖所示，用一條直線明顯可分。

綫性分類器的局限性

Q：另外一個例子：第一類：1，2，3　第二類：-3，-2，-1，4，5，6

空間變換，一維空間→二維空間，即 x → (x，x^2)

變換後數據為：第一類：(1,1)(2,4)(3,9)　第二類：(-3,9)(-2,4)(-1,1)(4,16)(5,25)(6,36)



如圖所示，用一條直線明顯可分。

SenseStudy實驗

SenseStudy課程平臺“人工智能入門（下）”： 實驗7 – 1 線性分類器-商鋪類型分類

```
coor_x, coor_y, label = load('eshop.train')
fig() + scatter(coor_x, coor_y, c=label)
model = LinearClassifier()
feat = model.merge_features(coor_x, coor_y)
model.train(feat,label)

coor_tx, coor_ty, t_label = load('eshop.test')
t_feat = model.merge_features(coor_tx, coor_ty)
pred = model.predict(t_feat)
accuracy = model.get_accuracy(t_feat, t_label)
print("the predicted accuracy is %f"%accuracy)

coor_x, coor_y, latent_v, label = load('tr_eshop.train')
feat = model.merge_features(coor_x, coor_y, latent_v)
model.train(feat, label)

coor_tx, coor_ty, tlatent_v, t_label = load('tr_eshop.test')
t_feat = model.merge_features(coor_tx, coor_ty, tlatent_v)
pred = model.predict(t_feat)
accuracy = model.get_accuracy(t_feat, t_label)
print("the predicted accuracy is %f"%accuracy)
```

課後作業

SenseStudy課程平臺“人工智能入門（上）”

實驗5 – 4 線性回歸模型的訓練演算法

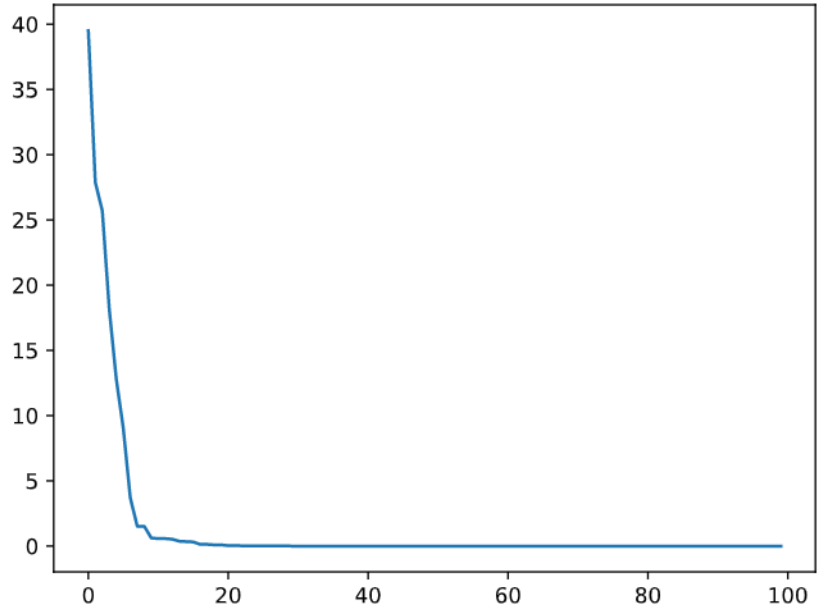
對於已有訓練集 x,y:

train_x = [0,2,4,6]

train_y = [0,4,8,12]

fig() + scatter(train_x, train_y)

使用反覆運算演算法求尋找y=ax 中的係數 a



課後作業

SenseStudy課程平臺“人工智能入門（上）”

實驗5 – 7 多變量分類模型

1、收集身邊某個應用場景下的數據

2、使用此實驗中介紹的綫性分類方法，訓練一個分類模型

3、使用此模型對更多數據進行分類，並評估分類結果

```
Predicted results are:  1 1 1 -1 1 1

Ground truth results are:  [1, 1, 1, -1, -1, -1]

Accuracy is:  0.6666666666666666
```