2016 年 3 月 15 日下午,舉世矚目的圍棋人機大戰在圍棋世界冠軍李世石與人工智慧博弈軟體 AlphaGo 之間展開第五局對戰。最終李世石在萬般無奈之下投子認輸,這場曠世圍棋人機大戰以 AlphaGo4:1 勝出的結局落下了帷幕。此後,AlphaGo 及其後繼者 AlphaZero 一路高歌猛進,所向披靡.橫掃世界圍棋職業棋壇,一次又一次向人類展示出其超強的人工智慧智力,而它的靈魂正是本書所關注的核心內容一一機器學習演算法。從 20 世紀 90 年代開始,互聯網改變了人們的生活。谷歌、亞馬遜等互聯網公司為資訊全球化做出了重大的貢獻。進入 21 世紀後,蘋果智慧手機 iPhone 的出現再一次帶來了技術的革新。諸多 App 層出不窮,為人們的衣食住行帶來了巨大的便利。下一輪創新的浪潮將來自何方?許多學者、工程師與企業家認為,人工智慧將引領未來的潮流。

人工智慧的概念是由以麥卡賽、明斯基、羅切斯特和香農等為首的一批科學家在 1956 年提出的。為什麼一個已有 60 餘年歷史的學科又重新走進了人們視野的中心?回顧歷史,每一次潮流的興起都伴隨著科技的創新。例如,互聯網的興起源於高速光纜的問世,手機 App 的繁榮源于智慧手機的誕生。那麼,又是什麼新興的技術突破為人工智慧領域注入了新鮮的活力?可以說,人工智慧的核心是機器學習,而機器學習的核心是演算法。近年來,機器學習演算法的研究屢屢取得重大成功,特別是深度學習演算法,更是一次次地展示出無與倫比的威力。同樣重要的是.GPU(影像處理器)的高速發展,使得大規模深度學習成為可能。所以,正是機器學習演算法理論及相應硬體技術的突破使人工智慧煥發新生。談到人工智慧

與機器學習,人們眼前會立刻浮現出諸多異彩紛呈的場景。在這些場景中,有無人駕駛汽車穿行於車水馬龍.有智慧型機器人探索宇宙太空,有面部識別系統精確定位尋蹤於茫茫人海,還有 AlphaGo 智慧博弈軟體與人類圍棋世界冠軍激戰,如圖 1.1 所示。所有這些場景都源於機器學習帶來的新一輪技術創新浪潮。十幾年前,這些場景也許只出現在科幻影視作品中,而今天,人工智慧的創新將它們變成了現實,並且徹底地改變了人們的生活。那麼,是什麼樣的技術支援著這些精彩紛呈的應用?機器如何通過學習獲取智

聚類分析(clustering analysis)是依據資料相似度或相異度而將資料分群歸屬到數個聚類(clusters)的方法;使得同一群內的資料或個體相似程度大,而各群之間的相似程度小。同一組樣本有時會因為不同的目的、資料登錄方式、所選的分群特徵或資料屬性,形成不同的分群結果。例如,圖 6.1(a)的資料,可以根據某些特徵和準則,將資料分成 3 個(圖 6.1(b))或 4 個(圖 6.1(c))聚類。另一方面,分類(classification)則是根據已知或所給定目標資料的類別,找出其分類屬性,建立分類規則或模式,將資料分類至所對應的目標類別。(c)4 個聚類 (a)原始資料點 (b)3 個聚類圖 6.1 不同的分群結果

聚類分析是分群以找出各子聚類資料背後可能隱藏的特徵、樣型或關聯現象。聚類分析事先並不知道聚類數目,而分群結果的特徵及其所代表的意義僅能事後加以解釋。因此,聚類分析可視為無監督式學習;而分類方法則視為監督式學習。聚類分析應用的領域相當廣泛。例如,根據顧客基本資料和交易資料將顧客分群,定義並分析不同類型顧客的消費行為模式,以設計定制化的行銷方案;或是通過聚類分析將信用卡使用行為分為不同群組樣型,以分析信用卡異常消費的情形,避免盜刷所造成的損失。在製造業,可依據機台的特徵、功能等的相似程度,將機台分為可以相互替代和備援(backup)的聚類,以提升作業效率並維持良率(Chien&Hsu,2006)。在網路行銷中,可將性質或特性相仿的網頁予以分類,增快網頁搜索速度,並根據流覽行為和客戶聚類分析作客戶消費行為預測和搭配行銷。此外,聚類分析也常常與其他演算法整合,將分群結果輸入後續的分析中。例如,提取各聚類的特徵作為後續分類的準則;或在資料準備時,運用聚類分析決定群組並將離散資料以代碼表示。

聚類分析主要包括以下四個階段:

(1)資料準備與分群特徵選取:根據問題特性、資料類型及所選擇的分群演算法等,自搜集的變數中選取具代表性的變數作為分群特徵屬性。

(2)相似度計算:選擇衡量相似度的方式,如距離、相關係數等。在選擇衡量相似度的方式時,需考慮資料的類型以及後續使用的分群演算法,例如,在類別尺度中,選用歐氏距離可能會造成資料尺度的誤用。

(3)分群演算法:為整個聚類分析中最重要的階段,主要為利用分群演算法將資料分組,有些分群演算法可能需要自行決定群數,例如,劃分聚類分析演算法可由使用者自行決定或利用其他方式決定適當的分群個數。

(4)分群結果評估與解釋:當分群結束後需檢查分群結果是否合理。例如,聚類間的距離是否過大、該資料是否適用所選用的分群演算法,若發現有不合理的地方,則需重新審視前三個階段是否有問題。另外,由於分群後的結果可能作為另一個方法的輸入資料,因此可能需要對聚類結果進行定義或命名。

由式(6.6)可知,

相關係數與單位無關;且相關係數介於-1 到+1 之間。

當 r(V1,V2)>0

表示 V1 增加時,V2 也增加;

r(V1,V2)<0 表示 V1 增加時,V2,則減少。

0<=|r(V1,V2)|<=0.3 表示兩變數為低相關性,

0.3<=|r(V1,V2)|<=0.7 表示兩變數為中相關性,

0.7<=|r(V1,V2)|<=1 表示兩變數為高相關性。

監督式學習的主要任務是根據物件的特徵來預測其標籤。相應的學習演算法需要對經驗進行學習。因此,訓練資料是監督式學習演算法的一個重要組成部分。訓練資料是隨機抽取觀測物件採集到的資料。這些資料簡稱為採樣,也稱為樣本。每一條訓練資料都含有特徵與標籤。例如,房價預測問題的訓練資料是過去一年的房屋成交記錄,它含有各種房屋的特徵以及售價。通過對訓練資料的學習,演算法能夠訓練出一個模型來預測標籤,並且根據給定的度量方法來檢驗模型預測的效果。

為了正式定義監督式學習,需要介紹以下幾個基本要素:特徵、標籤、分佈、模型與損失函數。下面逐一給出它們的正式定義。

定義 2.1(特徵組) 在一個監督式學習問題中,將每個物件的幾個特徵構成的向量 $X = (x1, x2, x3, \dots, xn) \in \mathbb{R}^n$稱為該物件的特徵組。設$X \subseteq \mathbb{R}^n$是特徵組的所有可能取值構成的集合,稱 X 為樣本空間。

例如,房價預測問題中,預測的物件為某社區的房屋。每個房屋都有房屋面積、臥室數和房齡 3 個特徵。這 3 個特徵構成了每個房屋的特徵組。如果某個房屋的面積為 100m,有兩間臥室,房齡為 15 年,則該房屋的特徵組 x=(100,2,15)。監督式

學習的每一條訓練資料都帶有標籤。根據標籤的不同形式,監督式學習又分為兩類:回歸問題與分類問題。一個含有 k 個類別的分類問題稱為 k 元分類問題。

定義 2.2(標籤) 在回歸問題中,訓練資料含有一個數值標籤$y \in \mathbb{R}$;在 k 元分類問題中,訓練資料含有一個向量標籤$y \in [0,1]^k$。設 Y 為全體可能的標籤取值,稱 Y 為標籤空間。為了解釋定義 2.2,首先探討回歸問題。回歸問題的標籤是物件的一個數值屬性,所以可以用一個數值$y \in \mathbb{R}$來表示標籤。例如,在房價預測問題中,房屋交易價格是房屋的一個數值標籤。

再來探討分類問題。分類問題的標籤表示物件的類別。在一個含有 k 個不同類別的分類問題中,每個物件的標籤都可以表示為一個 k 維向量。如果該物件屬於第 t 類,則標籤向量的第 t 位的值是 1,其餘位的值是 0。所以,標籤的可能取值是全體 k 維 0-1 向量$\{0,1\}^k$。例如,在手寫數字識別問題中,標籤是一個 10 維向量。如果訓練資料圖片中是數位 3,則標籤為(0,0,0,1,0,0,0,0,0,0)。為了便於敘述,通常定義分類問題的標籤空間為$[0,1]^k$,即分量均屬於[0,1]區間的 k 維向量全體。

在監督式學習中,有兩個重要的假設:

(1)物件(或特徵組)不是無規律出現,而是服從一定的概率分佈。

(2)給定對象的標籤也不是無規律生成的,而是服從某個由物件特徵組決定的概率分佈。

定義 2.3(特徵分佈) 設 X 為樣本空間。監督式學習假定特徵組是由一個 X 上的概率分佈 D 生成的。稱 D 為特徵分佈。用 x～D 表示 x 為依特徵分佈 D 的一個隨機採樣。

定義 2.4(標籤分佈) 設 Y 為標籤空間。對任意的$x \in X$,監督式學習假定存在一個由 x 決定的 Y 上的概率分佈 D,使得 x 對應的標籤 y 服從 Dx。稱 Dx 為特徵組 x 的標籤分佈。用 y～D 表示 y 为依分布 D・随机生成的标签。例如,在房價預測問題中,特徵分佈是正比於人口分佈的。在人口越密集的區域,房屋交易越頻繁,產生樣本的概率也越高。對任意給定的房屋特徵,房價也是一個隨機變數。對一批十分相似的房屋,房價未必完全相同。它可能是以某個價格為期望的正態分佈,這就是標籤分佈。當然在有的問題中,標籤由特徵唯一確定。例如,在手寫數位識別問題中,大多數情況下標籤值是唯一確定的。此時,標籤分佈退化為一個單點分佈。

有了標籤分佈的概念,就可以嚴格定義監督式學習的任務了。前面曾將監督式學習的任務概要地描述為根據特徵對標籤做出預測。實際上可以更加精確地描述如下:對任意的特徵組$x \in X$,如果將標籤 y 看作分佈為 D,的隨機變數,則監督式學

習演算法的任務是預測 y 的期望值$E_{y\sim D_x}$ [y]。無論是回歸問題還是分類問題,都有$E_{y\sim D_x}$ [y] ∈ Y,因此可以認為監督式學習的任務是計算從樣本空間 X 到標籤空間 Y 的映射。這樣的映射就稱為一個模型。

定義 2.5(模型) 設 X 為樣本空間,Y 為標籤空間,Φ 為全體 X Y 的映射集合。稱 4 為模型空間。任意模型空間中的映射 h 中都稱為一個模型。

綜合定義 2.1 至定義 2.5,可以對監督式學習任務正式定義如下。

在一個監督式學習問題中,給定樣本空間 X、標籤空間 定義 2.6(監督式學習任務) 在 Y、未知的特徵分佈 D 與標籤分佈 DEX,監督式學習任務是計算一個模型 hX+Y,並對任意特徵組 x,以 h)作為對標籤期望值 E,~D[y]的預測。也將 hx)簡稱為對 x 的標籤的預測。

以下介紹對監督式學習演算法效果的度量方法。用模型 h 對 x 的標籤值做預測時,預測結果與真實情況可能存在誤差。為了度量這個誤差的大小,需要引入損失函數的定義。

[範例 6.1]為 7 筆觀察值的 V1 與 V2 資料(如表 6.2),為方便計算,以歐式距離平方作為衡量相似度的依據,可計算出各資料點間的歐式距離平方如表 6.3 所列。假設現在有三個聚類,分別是聚類 A={1,3,6},聚類 B={2,4},聚類 C={5,7},聚類 A 與 B 間

共有 6 個距離,分別為:$D_{1\&2}$=233、$D_{1\&4}$=261、$D_{3\&2}$=149、$D_{3\&4}$=169、$D_{6\&2}$=80、$D_{6\&4}$=104.

表 6.2[範例 6.1]觀察值

觀察值	V1	V2
Y1	14	15
Y2	22	28
Y3	15	18
Y4	20	30
Y5	30	35
Y6	18	20
Y7	32	30

歐氏距離平方

序號	1	2	3	4	5	6	7
1	0	233	10	261	656	41	549
2	233	0	149	8	113	80	104
3	10	149	0	169	514	13	433
4	261	8	169	0	125	104	144
5	656	113	514	125	0	369	29

6	41	80	13	104	369	0	296
7	549	104	433	144	29	296	0

最小距離： $D_{min(C_A,C_B)}=D_{6\&2}=80$

最大距離： $D_{max(C_A,C_B)}=D_{1\&4}=261$

平均距離： $D_{average(C_A,C_B)}=\frac{D_{1\&2}+D_{1\&4}+D_{3\&2}+D_{3\&4}+D_{6\&2}+D_{6\&4}}{6}=166$

中心值距離： *聚類A的中心*：$\left(\frac{14+15+18}{3}\quad\frac{15+18+20}{3}\right)=\left(\frac{47}{3}\quad\frac{53}{3}\right)$

聚類B的中心：$\left(\frac{22+20}{2}\quad\frac{28+30}{2}\right)=(21,29)$

則，聚類 A 與聚類 B 的歐氏距離為：

$$D_{centroid(C_A,C_B)}=\left(21-\frac{47}{3}\right)^2+\left(29-\frac{53}{3}\right)^2=156.89$$

常見的層次聚類分析方法包括:單一連結法(single linkage method),以兩聚類間資料點中的最小距離來表示兩聚類的距離及兩群資料的鄰近程度;完全連結法(complete linkage method),以兩聚類間資料點的最大距離來表示兩聚類的距離及兩群資料的鄰近程度;平均連結法(average linkage method),衡量聚類內所有點到另一個聚類內所有點的距離平均來表示兩聚類的鄰近程度,以避免聚類之間的距離衡量受雜訊影響;中心點連結法(centroid linkage method),以兩聚類的中心點距離作為衡量兩聚類的距離,以表示其鄰近程度。

以[範例 6.1]為例,利用單一連結法說明層次聚類分析的計算,起初所有資料皆屬於單一聚類,而資料點 2 與資料點 4 最接近,所以將兩點合併為一聚類,重新計算各聚類間資料點的最小距離如表 6.4 所示。而資料點 1 與資料點 3 最為接近,因此將兩點合併為新的聚類，迭代，直到將所有數據點均合併至同一聚類中為止。

單一連結法，合併 2 和 4 後的歐氏距離

序號	1	2&4	3	5	6	7
1	0	233	10	656	41	549
2&4	233	0	149	113	80	104
3	10	149	0	514	13	433
5	656	113	514	0	369	29
6	41	80	13	369	0	296
7	549	104	433	29	296	0

最後聚類 AB 與聚類 C 在距離為 104 時合併為一群

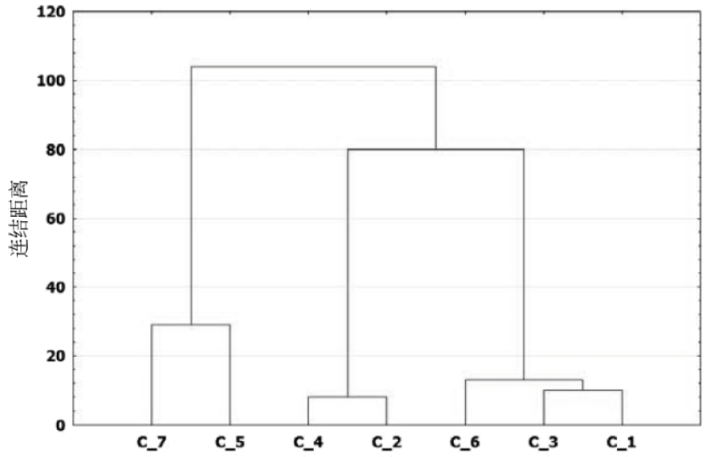

图 6.4　单一连结法树形图

K-means

表 6.6　起始聚类中心

聚类	V_1	V_2
A	14	15
B	20	30
C	18	20

表 6.7　K 平均法分群过程(初始重新分配)

序号	与聚类中心的距离			最小距离	分配的聚类
	聚类 A	聚类 B	聚类 C		
1	0	261	41	0	A
2	233	8	80	8	B
3	10	169	13	10	A
4	261	0	104	0	B
5	656	125	369	125	B
6	41	104	0	0	C
7	549	144	296	144	B

表 6.8　聚类中心(第一次重新分配)

聚类	V_1	V_2
A	14.5	16.5
B	26	30.75
C	18	20

表 6.9　*K* 平均法分群过程(第一次重新分配)

序号	与聚类中心的距离			最小距离	分配的聚类
	聚类 *A*	聚类 *B*	聚类 *C*		
1	2.5	392.06	41	2.5	*A*
2	188.5	23.56	80	23.56	*B*
3	2.5	283.56	13	2.5	*A*
4	212.5	36.56	104	36.56	*B*
5	582.5	34.06	369	34.06	*B*
6	24.5	179.56	0	0	*C*
7	488.5	36.56	296	36.56	*B*

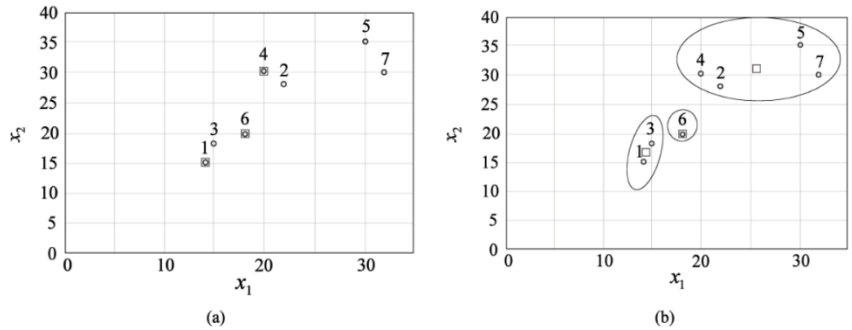

图 6.6　*K* 平均法对[范例 6.1]的分群过程

然而,若實際的分群結果為聚類 A=(1,3,6),聚類 B=(2,4),聚類 C=(5,7),總距離變異平方和會較上面的結果小。由此可知,選擇起始資料作為聚類中心可能會影響分群的結果。

K 平均法雖然已廣泛使用在聚類分析上,但仍存在一些缺點:

・K 平均法无法直接处理类别型的数据(因无法求得数据的中心点),这类型数据可改用另一種劃分聚類分析法 K 眾數法(K-mode)進行分群。K 眾數法是用簡單配對相異度(simple matching dissimilarity)作為衡量相似度的指標,並以聚類的眾數作為聚類的中心,用頻率為基礎(frequency-based)的方法來更新聚類的眾數,詳細內容可進一步參見(Huang,1998)。

・K 平均法必須事先決定聚類數目。聚類數目往往需由使用者直接給定,或通過反復分析與驗證,取得適當的群數。另外可利用兩階段的方式,也就是先用層次聚類分析演算法決定聚類的數目,再利用 K 平均法重新將資料歸類分群(Sharma,1996)。

・分群结果容易受到离群值的影响。因為 K 平均法是以平均值作為聚類的中心,在計算時容易受到離群值的影響造成偏移,產生聚類分佈上的誤差。為了避免離群值影響分群結果,可改用 K 中心點法進行分群。

・起始聚类中心选择的不同会造成不同的分群结果,若起始聚类中心的数据不够分散,可能会得到较差的聚类结果。

・无法适用于所有的数据聚类形态,如 K 平均法无法处理非球状的聚类、数据大小差异很大的聚类,和数据密度不同的聚类。

當聚類間的特性非常相似時,在邊界上的資料點只要有一點偏差,就可能從 A 聚類劃分到 B 聚類。這類型的資料可改用柔性聚類(soft clustering)方法來處理。