

心理實驗法第 16 週作業

公衛三 梁嫚芳 B07801003

使用 JASP 分析 d1.0.csv 來比較統計推論與機器學習。

1. 使用獨立樣本 T-test: 報告顯著性 p, Cohen's d, & BF10 (須截圖，各 1 分)

$p < .001$

Cohen's d = -0.955 (95% CI: (-1.047, -0.862))

BF10 = 2.928e+87

Independent Samples T-Test ▼						
	t	df	p	Cohen's d	95% CI for Cohen's d	
					Lower	Upper
X	-21.351	1998	< .001	-0.955	-1.047	-0.862

Note. Student's t-test.

Bayesian Independent Samples T-Test		
	BF ₁₀	error %
X	2.928e +87	2.532e -90

2. 使用 kNN 分類: 報告 Validation & Test Accuracy (須截圖，各 1 分)

Validation Accuracy = 0.688

Test Accuracy = 0.620

K-Nearest Neighbors Classification							
Nearest neighbors	Weights	Distance	n(Train)	n(Validation)	n(Test)	Validation Accuracy	Test Accuracy
9	rectangular	Euclidean	1280	320	400	0.684	0.623

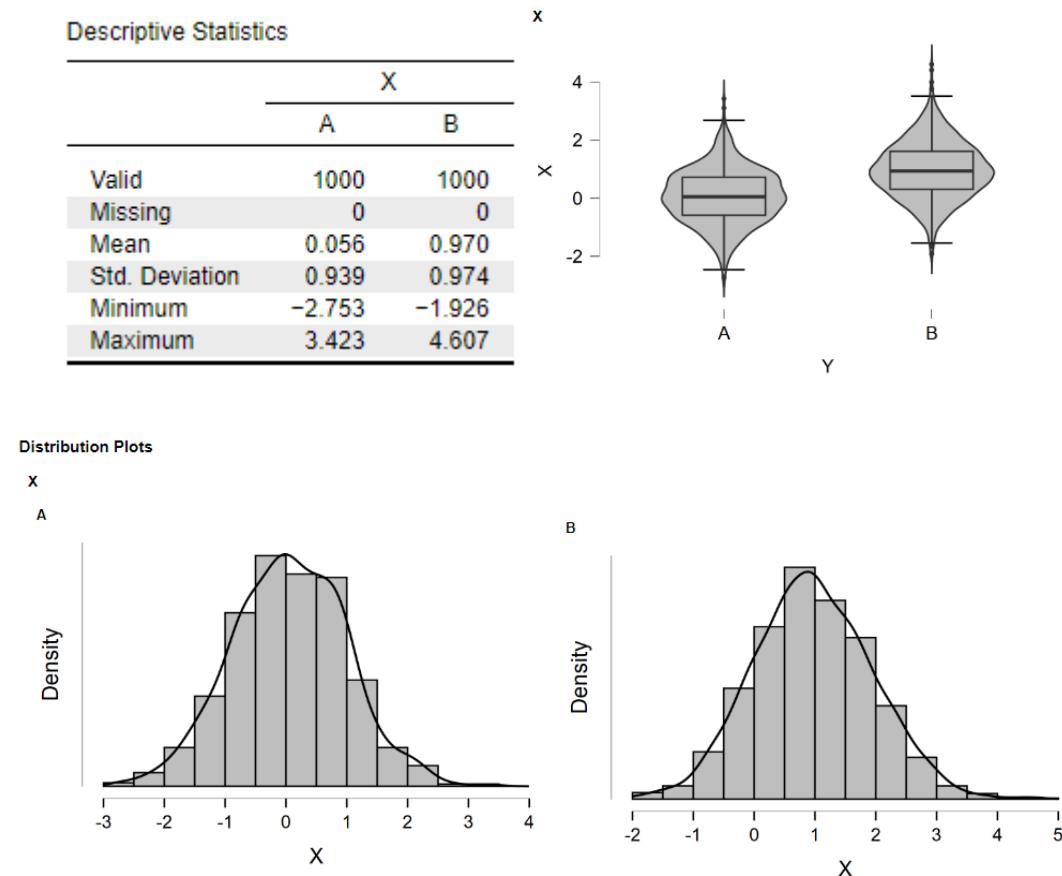
Note. The model is optimized with respect to the validation set accuracy.

Data Split

Train: 1280	Validation: 320	Test: 400	Total: 2000
-------------	-----------------	-----------	-------------

3. 討論為何統計效應這麼顯著且效應這麼大，機器學習卻沒辦法預測很好? (2分)

以描述性統計表格來看，A 與 B 組之 X 的平均值差異接近 1，標準差數值皆近似 1，以下方 violin plot 來看，B 組之 X 稍高於 A 組，接著，觀察下方 distribution plot，可發現 A 與 B 組之 X 分布極大部分都重疊。

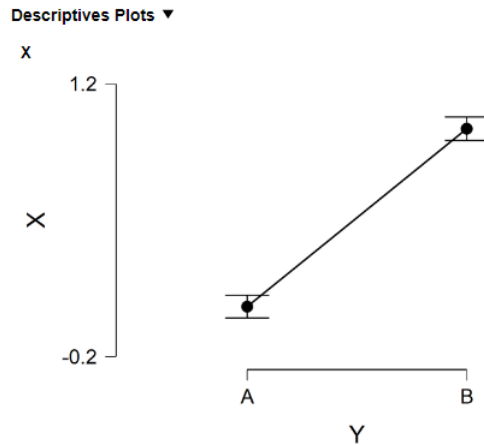


以統計檢定來看，下圖可見 A 與 B 組之 X 平均值的信賴區間均十分窄，且距離差異顯著，為何於前段描述性統計時，乍看兩分布大多重疊，卻會如此顯著差異呢，我認為此與統計的定義有關。

針對信賴區間，以 t 分布來說，公式為：

$$\bar{x} \pm t_{\alpha/2, v} \times \frac{s}{\sqrt{n}}$$

t 為事先決定之定值(如 95% CI 約為 1.96)，因此可知 S (標準差)愈小，n (樣本數)愈大，信賴區間寬度愈小，而此資料特徵為 s 不大 (0.939, 0.974)，且 n 極大，A 與 B 組之 n=1000，因此 s 除以 $\sqrt{n} \approx 31.6$ 後，信賴區間即縮小得極窄，彌補 mean X 相差不遠 (約為 1) 的較難顯著的劣勢，從而促成倆距離遙遠的信賴區間，並於 t 檢定時顯著(p < .001)。



針對 Cohen's d，公式為：

$$d = \frac{\bar{x}_A - \bar{x}_B}{S_{pooled}}$$

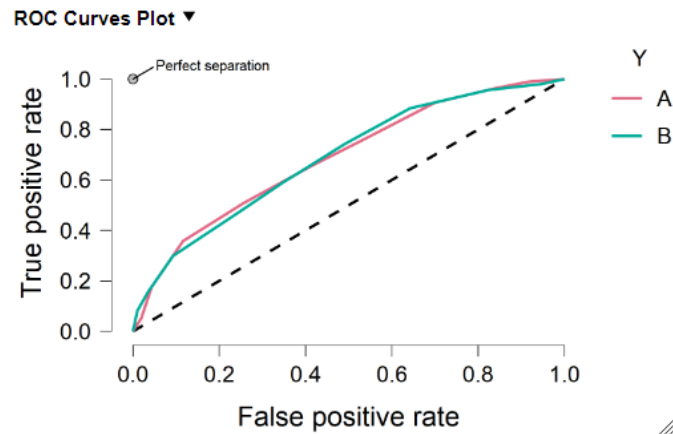
其中， \bar{x}_A 與 \bar{x}_B 分別為兩組之 X 平均值， S_{pooled} 公式如下：

$$S_{pooled} = \sqrt{\frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{n_A + n_B - 2}}$$

因此可見，若要彌補兩組 X 平均值差異僅接近 0.913，來達成如此大的 Cohen's d， S_{pooled} 勢必要小到接近平均值差異，這點可由標準差不大來達成，A 與 B 組標準差分別為 0.939, 0.974，僅較平均值差異較大一些，因此 Cohen's d 最終計算結果為 $-0.955 > 0.8$ 非常大。

針對 BF10，若樣本數 n 愈大，則 Bayes factor 愈大，而此資料之樣本數極大，經由 Bayes Factor for Grouped or Two-Sample t-Tests (<http://pcl.missouri.edu/bf-two-sample>) 網路計算機測試結果具有顯著差異，當 n=10 時 BS10 約為 $9.1e+10$ ，當 n=1000 時 (此資料) BS10 約為 $2.9e+87$ ，遠大於 100。

以機器學習而言，KNN 為以鄰近樣本來判斷屬於哪個群集，能夠預測得最佳的情況為，A 與 B 組的 X 分布得足夠離散 (分成兩群)，或者數個群集 (clusters)，而非混雜，然而此資料中，兩組之 X 幾乎重疊，因此預測時表現不佳。儘管前述兩分布於統計檢定時差異顯著，為統計意義上兩分布間有差異，但由於兩分布幾乎交織在一起，中間重疊的部分於預測時就無法透過此差異去進行良好的判斷。



由機器學習模型的 ROC curve 可見，兩條 AB 曲線非常靠近虛線，代表此分類工具無論敏感性 (Sensitivity; Recall) 或特異度 (Specificity) 均不佳，以篩檢工具而言，此即為兩分布重疊較大的結果。

我認為統計是對數據的數學研究與推論解釋，機器學習則犧牲可解釋性以獲得強大的預測能力，目的不同，因此兩者間結果的可比性不高，正如此題中的情況，雖於統計檢定中取得顯著，機器學習卻無法預測得宜。