實作心得報告

NICE: an algorithm for nearest instance counterfactual explanations

簡述 NICE:

NICE 演算法的核心目標在於提供對分類結果的反事實解釋，這意味著在保持其他 features 變動最少的情況下，揭示出將樣本分類至不同類別的可能性。這種能力不僅有助於使用者深入了解機器學習模型的判斷過程，更提供了一個實際的指引，讓使用者了解在什麼條件下可以改善或調整以達到更理想的結果，此演算法的特點在於因為高速，且保證找到 counterfactual。

使用申請貸款舉例，要成功申請貸款，或許需要提高月收入這一屬性。

實作方法:

我們的實作方式，是使用其他 dataset 觀察輸出結果，我們使用了 kaggle 上面取得的資料集: Stroke Prediction Dataset (kaggle.com) 。

此 dataset 展示了年齡、血糖、高血壓等與中風息息相關的徵兆，

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	1	67.0	0	1	1	2	1	228.69	36.600000	1	1
1	0	61.0	0	0	1	3	0	202.21	28.893237	2	1
2	1	80.0	0	1	1	2	0	105.92	32.500000	2	1
3	0	49.0	0	0	1	2	1	171.23	34.400000	3	1
4	0	79.0	1	0	1	3	0	174.12	24.000000	2	1
...	...	...	...	...	...	...	...	...	...	...	...
5105	0	80.0	1	0	1	2	1	83.75	28.893237	2	0
5106	0	81.0	0	0	1	3	1	125.20	40.000000	2	0
5107	0	35.0	0	0	1	3	0	82.99	30.600000	2	0
5108	1	51.0	0	0	1	2	0	166.29	25.600000	1	0
5109	0	44.0	0	0	1	0	1	85.28	26.200000	0	0

5110 rows × 11 columns

以及諸如工作，居住地點，婚姻狀態等隱性原因的類別資料。

目前的 NICE 演算法不夠全面，所以只能使用二元分類結果的資料集，例如上述的預測有沒有中風。

而 NICE 演算法支援使用三種不同的 unlike neighbor 的方法，分別為 Sparsity(只使用較少的 features),Proximity(nearest unlike neighbor),Plausibility(結果需要符合現實邏輯)，本次使用的是最基本的 proximity。

得到以下結果:

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	0.0	40.0	0.0	0.0	0.0	2.0	0.0	158.93	31.3	3.0	0
1	1.0	49.0	0.0	0.0	0.0	2.0	0.0	104.86	31.9	3.0	1

此結果中第 0 行為起點，而第一行為 NICE 找到的 nearest unlike neighbor, 如此透過此例子可以看到，血糖濃度雖下降嚴重，但是年齡的提升能讓此患者患上中風，而兩個的性別也有所不同(0:male, 1:female)，在後續會做相關的討

論，我們能透過此解釋概略地觀察模型，在此狀況中此模型將 Age 的權限提高了不少，

我們在觀察其他的 instance。

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
0	1.0	47.0	1.0	0.0	1.0	2.0	1.0	75.64	24.4	2.0
1	0.0	51.0	1.0	0.0	1.0	2.0	1.0	75.64	28.4	2.0

得到的結果也與年齡較為相關，可以說年齡確實為主要因素，同時經由和一開始的例子做比較，可以看出年齡不管是男變女，還是女變男，都很難探討出對中風的影響，或許是個較無重要的因素。

心得:

經過實作，感覺到目前演算法還是有著不少的侷限性，但是其演算法非常明瞭簡潔，所以相信擴充性應該是非常良好的。