工人智慧到人工智慧：
大數據與文字探勘在法律經濟學的應用
--以商標侵權為例

成功大學 林常青 張芝嘉

中興大學應用經濟系
2021/04/30

法學研究的典範類型

<table>
<tr><th rowspan="3">研究主題:
法律</th><th colspan="4">分析方法</th></tr>
<tr><td rowspan="2">規範分析
(normative)</td><td colspan="3">實然分析 (positive analysis)</td></tr>
<tr><td>不用資料</td><td colspan="2">應用資料</td></tr>
<tr><td>非使用社會科學方法</td><td>法釋義學</td><td>非實證的實然法學</td><td>實證法學</td><td rowspan="3">合稱:
法律實證研究</td></tr>
<tr><td rowspan="2">使用社會科學方法</td><td>規範社科法學 (如: 政策、福利經濟學、羅爾斯正議論)</td><td>非實證的實然社科法學 (經濟, 政治, 社會理論)</td><td>實證社科法學</td></tr>
<tr><td colspan="3">合稱: 社科法學 (如 法律經濟學, 法律社會學)</td></tr>
</table>

張永健, 2019，〈第一章: 法實證研究的方法座標〉，《法實證研究: 原理、方法與應用》，表2, 頁16。

實然分析的社科法學的主要類型與例子

社科法學	實然分析	
	不用資料	應用資料
法律經濟學	數理法律經濟學	實證法律經濟學 (成本、效益、動機)
法律社會學	法的社會分析 (馬克思、偉伯...) 理論	法律與社會研究 (法律與社會的互動)
法認知心理學	數理法律行為經濟學(如：偏見、捷思)	行為法經濟學、法認知科學
法政治學	公共選擇理論	司法政治
法人類學		定性或定量法人類學

張永健, 2019，〈第一章: 法實證研究的方法座標〉，《法實證研究: 原理、方法與應用》，表3, 頁17。

實證法學　v. 實證社科法學

- 實證法學:
  - 以為法學自身的關心為主: 如
    - 美國政府徵收時總是給予財產所有人過少補償,
    - 民眾偏好認為有罪當無罪(錯放)或是無罪當有罪(錯判)哪個嚴重?

- 實證社科法學:
  - 問題意識以社科法學的理論或觀察出發, 研究法律相關題材
  - 一般性, 能被其他社科法接受 如:
    - 定錨效應,
    - 人際關係...

法律實證研究的四種作用:

1. 規範論證的實然基礎
   - 應然論證 (1 大前提) A是好的 (2 小前提) B 可達 A (3 結論) 應採 B
   - 結果論證 (1 大前提) B 可達 A (2 小前提) A是好的 (3 結論) 應採 B

- 法律人關心 (評價) A是好的 以及 (應然) 應採 B
- 法律實證關心 (實然) **B 可達 A ??**

張永健, 2019，〈第一章: 法實證研究的方法座標〉，《法實證研究: 原理、方法與應用》，頁27-32。

法律實證研究的四種作用:

2. 法律的實效性??
3. 更有效地描述法律論證與法現象?
   - 法律適用的實際情形 → 新的問題意識, 規範分析的實然基礎
4. 發掘法制度相關行動者的行為模式
   - 實證法律經濟學、法律與社會研究、行為法經濟學、法認知科學…

張永健, 2019，〈第一章: 法實證研究的方法座標〉，《法實證研究: 原理、方法與應用》，頁27-32。

- 研究動機
- 資料
- 研究方法
- 估計結果
- 結論

◻ 研究動機——研究背景

- 商標是財產法中常討論的對象--其經濟理論基礎在於減少搜尋商品與認定品質。(Cooter & Ulen, 2012 Ch5)

- 商標法第71條(100.06.29)：擇一計算損害賠償金額。

一、具體損害計算說或差額說。

二、總利益說及總價額說。

三、零售單價多倍賠償說 (就查獲侵害商標權商品計算)。

就查獲侵害商標權商品之零售單價一千五百倍以下之金額。但所查獲商品超過一千五百件時，以其總價定賠償金額。

四、授權金額決定說。

(經濟部智慧財產局 (2013)、陳麗珣 (2016))

◻ 研究動機——文獻回顧

- 於本國判決之實務上，商標法之侵權態樣經常以僅僅**6項**分類作為概括損害賠償審酌因素。(經濟部智慧財產局， 2012；高愷均， 2015；智慧財產法院98年度至104年度共24個判決)
- **太過粗略、實際參考之審酌因素多有出入**

- 陳姵君和劉湘宜 (2013) ：**8項**　　陳麗珣 (2016)：整理出**12項**

- **個案分析、各因素交互干擾，常見審酌因素混亂**
- **樣本太少 (不超過207份)**
- **過去並未有以數量方法對大量的判決書進行商標權案統計分析之文獻**

□ 研究動機——文獻回顧

- 以電腦程式輔助處理大量判決。
- **量刑預測**：林琬真等(2012) 以強盜及恐嚇取財之裁判為例，使用CRF資料系統進行法律要素標籤自動標記將案件分類。(刑法：強盜及恐嚇取財)
- **親權裁判**：黃詩淳、邵軒磊 (2018) 分析出法官在考量中最具有影養力之因素，提高之可預測性及法之安定性。(民法：親權)
- **著作權**：林筱瓴 (2013) 以Perl語言撰寫設計資料庫知識發掘系統(Knowledge-Discovery in Databases, KDD) 進行資料探勘，統計判決中影響損害賠償金額之因子。(民法：著作權)
- **專利品質評估**：羅郁仁 (2011) 利用斷詞方法將關鍵指標建立專利品質評估類神經模型。(民法：專利)

- 未有對**商標**之大量案件，以電腦程式輔助，進行數量分析。

◻ 研究動機

- ✓ 利用計量方法，整理歸納**實務上**法院判定勝敗與損害賠償金額之**審酌因素與結果**。
- ✓ 利用斷詞結果，找出其他**非學界歸納出**影響法官勝敗判決結果之案件**特徵**。
- ✓ **人工語言處理(斷詞 + Word2Vec)+分群(SVM)**是否能較計量方法提升勝敗準確率?
- ✓ **其他方式處理?(XGBoots & 資料來源適當分類)**

◻ 資料——來源

- 政府公開資料庫之判決書：

◆時間：從民國93年至107年4月。

◆範圍：包含所有地方法院與智慧財產法院。

◆裁判類別：民事、判決、一般案件。

◆搜尋關鍵字：

  ◆「商標法63」：93年至101年7月1日 (新法條正式施行)。

  ◆「商標法71」：101年7月1日至107年4月30日。

◆判決書數量：449 件（智院：272件、地院：177件）。

◻ 資料——審酌條件

常見審酌因素 (陳麗珣，2016)	本研究整理後量化
加害人之事業經營規模	資本額
對市場交易秩序之衝擊	交易秩序
系爭商標之知名度	著名商標
實際查獲系仿品數量	查獲數 查獲是否巨量
原告所受之損害	銷售數量
侵害期間	侵權時間 時間是否過長
被告侵害系爭商標權所得利益	營業收入
侵權情狀 →	概括，要想辦法量化

資料——審酌條件

常見審酌因素 (陳麗珣，2016)	本研究整理後量化
故意或過失	故意 過失
刑事不起訴	刑附民
原告願意和解	
仿冒商標之相同或近似程度	混淆誤認
	善意及合理使用
	著作權法
	公平法
	民法
	其他非屬法學界認定之法官審酌因素??

僅3件，但屬於常見之案件分類

僅6件

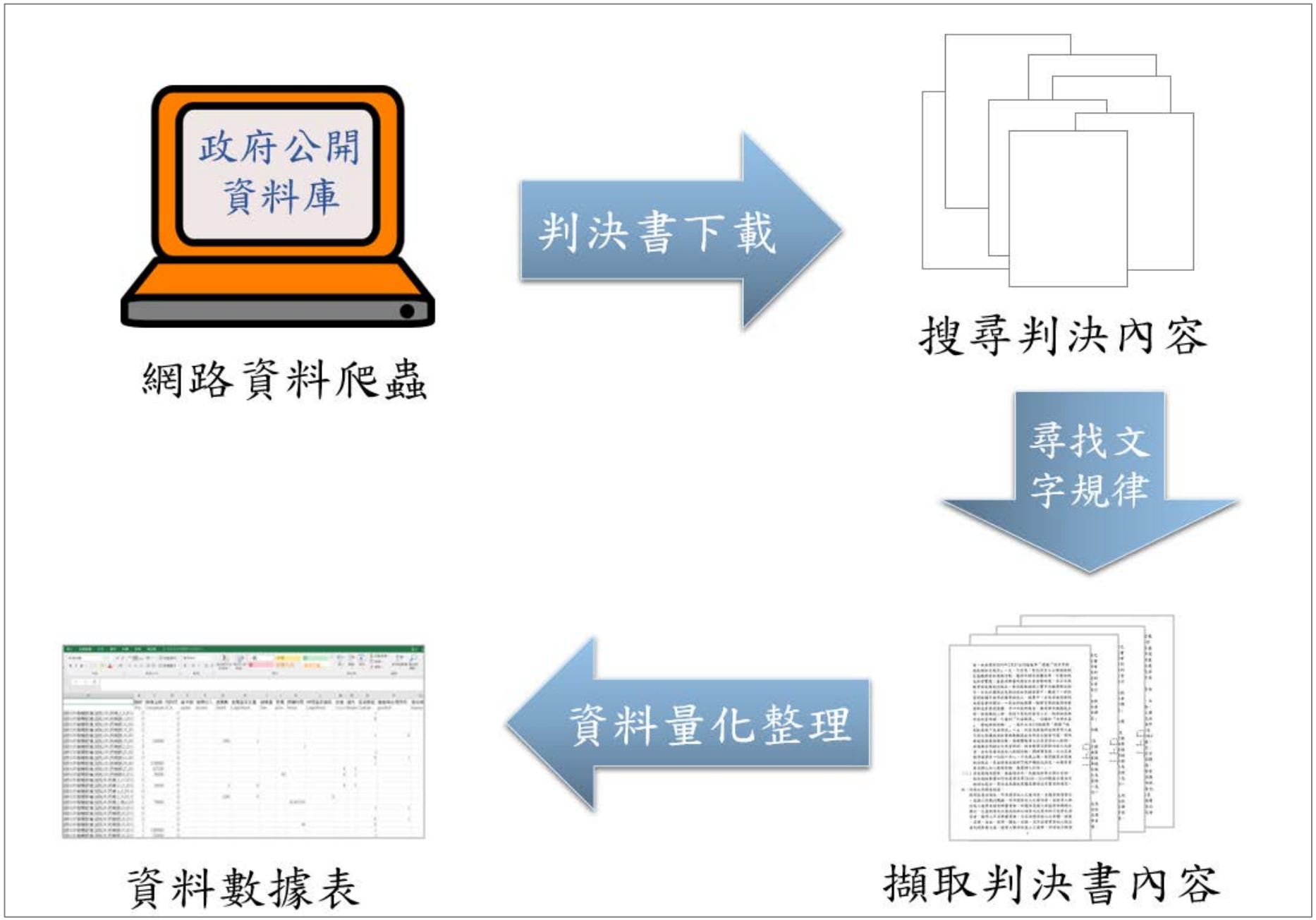
(80件)

◻ 資料——審酌條件

	經濟部智慧財產局 (2012)	陳姵君和劉湘宜 (2013)	陳麗珣 (2016)	本研究重新整理
加害人之經營規模	✓	✓	✓	資本額
仿冒商標商品之數量	✓	✓	✓	查獲數/查獲是否巨量
侵害行為之期間	✓	✓	✓	侵權時間/時間是否過長
仿冒商標之相同或近似程度」	✓	✓	✓	混淆誤認
在市場上流通情形	✓	✓	✓	著名商標
對市場交易秩序之衝擊			✓	交易秩序
被告侵害系爭商標權所得利益		✓	✓	營業收入
故意或過失			✓	故意/過失
刑事不起訴			✓	刑附民
原告所受之損害	✓		✓	

◻ 資料——審酌條件

	經濟部智慧財產局 (2012)	陳姵君和劉湘宜 (2013)	陳麗珣 (2016)	本研究重新整理
侵權商品金額		✓		相似
侵權商品單價		✓		
侵權情狀			✓	銷售數量
後續配合交代仿品來源		✓		(案例太少)
原告願意和解			✓	(案例太少)
				善意及合理使用
				著作權法
				公平法
				民法
				非屬法學界認定之法官審酌因素



本研究於Python程式語言使用流程圖

◻ 資料——判決內容處理

- 法官判決之理由(心證) 起始詞

1. 心證之理由
2. 心證理由
3. 本院之判斷
4. 本院判斷
5. 兩造之爭執
6. 兩造爭執
7. 本件爭點
8. 兩造主要爭點
9. 兩造均不爭執
10. 兩造不爭執
11. 兩造所不爭
12. 審酌

法官判決之理由結尾詞

1. 據上論結(斷,述)
2. 據上
3. 結論
4. 綜上所述
5. 綜上
6. 書記官

◻ 資料—判決內容處理

1. 相似的判斷字句在法律條文中重複出現，或參雜其他法條。如: 商標權損害賠償
   - 「按除本法第30條另有規定外，未經商標權人同意，而於同一商品或服務，或類似之商品或服務，使用相同於其註冊商標之商標者，有致相關消費者混淆誤認之虞者，其為侵害商標權。」
   - 公平交易法第25條規定：「除本法另有規定外，事業亦不得為其他足以影響交易秩序之欺罔及顯失公平之行為」
2. 重覆敘述被告或原告之主張。
   - 「主張原告過失侵害」
   - 「主張被告之專利權侵害行為係屬故意」
3. 中文同義詞複雜，且數字表達方式未統一。
   - 「獲利十四萬二千八百元」「獲得利潤大約有2,244,900元」「新臺幣伍拾伍萬元」

◻ 資料——判決內容處理

- 舉例：是否造成混淆誤認 (以商標法為例)

1. 相似之判斷字句 (同意義之詞句太多元)：

- 近似程度甚高、相關消費者難以區辨、與系爭商標圖樣相同.....等等。
- 相關消費者亦不會誤認、應屬構成近似程度甚低、故相關消費者極易辨識、已足以避免消費者混淆.....等等。

2. 去除非法官判定，為干擾項 (如：復述法條內容也可能有稍微刪改)：

- 使用相同於其註冊商標之商標者，有致相關消費者混淆誤認之虞 (商標法68)、使用相同或近似之商標，有致減損該商標之識別性(商標法70)、於同一或類似之服務為相同或近似之使用，致與他人營業或服務之設施或活動混淆者(公平法22) .....等等。

3. 若內容遇見「混淆」，判斷前置或後置有無否定 (有否定詞不一定是針對關鍵字，且須處理雙重否定)。

◻ 斷詞系統

◆英文斷詞相當容易，且詞性明確。

e.g. An apple a day keeps the doctor away

◆而中文斷詞有相當大的難度，無法明確斷詞，**需要建詞表**

e.g. 一個人才能睡覺

- A斷詞系統：一(Neu)/個(Nf)/人才(Na)/能(D)/睡覺(VA)
- B斷詞系統：一(Neu)/個(Nf)/人(Na)/才(Da)/能(D)/睡覺(VA)

◻ 詞頻

本文計算方式：

該詞出現的次數 = 若是文章中有出現，則計次一次

頻率 = 該詞出現的次數/總文章數

倍率 = A類文章詞頻/ B類文章詞頻

上訴人 Na 起訴 VC 起
務 Na 稱 Na 月泉 Na
業務 Na 期間 Na 無端
表 Na 示 VE 著名 VH
Na 損害 VC 損害 VC

判決書內文進行斷詞

辭彙整理

建立字典並做勝敗訴分類

計算詞頻

詞	勝-頻率	敗-頻率
上訴人	0.131403	0.160356
起訴	0.570156	0.35412
主張	0.576837	0.376392
侵害	0.565702	0.360802
著名	0.244989	0.167038
商標	0.563474	0.363029

整理歸納出現詞頻高者納入迴歸模型分析

本研究使用Python程式語言進行詞頻計算流程圖

□ 資料

- 審酌條件之勝率

	案件	勝(賠償)	勝率
全部	**449**	**272**	**0.606**
1. 刑附民	62	47	0.758
民事	387	225	0.581
2. 資本額	14	10	0.714
3. 查獲數	121	85	0.702
4. 查獲是否巨量	44	30	0.682
5. 銷售數量	6	5	0.833
6. 營業收入	8	8	1
7. 侵權時間	54	37	0.685
8. 時間是否過長	14	6	0.429
9. 交易秩序	71		
影響	47	25	0.532
不影響	24	15	0.625
10. 著名商標	82		
著名	65	44	0.677
非著名	17	5	0.294
11. 故意	116		
故意	75	57	0.76
非故意	41	24	0.585

資料

- 審酌條件之勝率

	案件	勝(賠償)	勝率
12. 過失	116		
過失	25	21	0.84
非過失	91	60	0.659
13. 善意合理	80		
善意	35	9	0.257
非善意	45	36	0.8
14. 混淆誤認	214		
混淆	133	95	0.714
不會混淆	81	21	0.259
15. 著作權法	47	34	0.723
16. 公平法	94	52	0.553
17. 民法	277	169	0.61

◻ 估計結果

1. 將判決書中是否提到前學術認定之17點審酌因素(以及4項特殊商品類別)加入解釋變數。
2. 利用逐步選取法 (stepwise selection) 進行篩選有統計顯著意義的自變項。

■Heckman兩階段估計法 (1976 –請見 Appendix)

■Exclusive variable (排除變數Z)：**估計結果顯著**

1. 善意與合理使用：商標法36條第1項第3款
2. 提到故意或過失：不具備故意或過失行為→不具有侵權事實

估計結果

	Heckman_twostep			
	賠償金額		勝訴	
刑附民	1.932	(1.297)	**0.511****	(0.190)
提到銷售量	1.039	(2.879)	0.077	(0.519)
故意	2.118+	(1.150)	**0.457***	(0.181)
過失	1.111	(1.737)	**0.734***	(0.308)
混淆誤認	-0.605	(1.005)	**0.328***	(0.154)
著名商標	-2.522*	(1.058)	0.105	(0.196)
民法	1.644*	(0.714)	-0.027	(0.128)
善意使用			-0.905***	(0.243)
截距項	12.546***	(2.377)	0.074	(0.115)
Inverse Mills ratio	-4.877+	(2.927)		
觀察值	449			

註：*表示統計上的顯著性：+為10%顯著水準，*為5%顯著水準，**為1%顯著水準，***為0.1%顯著水準。

詞頻比

- 敗訴

	A	B	C	D	E
1		次數	頻率	勝-頻率	敗/勝率
2	發酵	6	0.033898	0	--
3	名聲	5	0.028249	0	--
4	試圖	5	0.028249	0	--
5	中藥	5	0.028249	0	--
6	萃取物	5	0.028249	0	--
7	電源	5	0.028249	0	--
8	內銷	5	0.028249	0	--
9	牌子	5	0.028249	0	--
10	發還	5	0.028249	0	--
11	孩子	5	0.028249	0	--
12	用到	5	0.028249	0	--
13	臚列	8	0.045198	0.003676	12.29379
14	起始	8	0.045198	0.003676	12.29379
15	人體用	7	0.039548	0.003676	10.75706
16	手寫	7	0.039548	0.003676	10.75706
17	製劑	6	0.033898	0.003676	9.220339
18	恢復	6	0.033898	0.003676	9.220339
19	新舊	10	0.056497	0.007353	7.683616
20	尋得	5	0.028249	0.003676	7.683616
21	擔心	5	0.028249	0.003676	7.683616
22	誘因	5	0.028249	0.003676	7.683616
23	疏失	5	0.028249	0.003676	7.683616
24	信義	5	0.028249	0.003676	7.683616
25	移轉給	5	0.028249	0.003676	7.683616
26	過來	5	0.028249	0.003676	7.683616
27	代購	5	0.028249	0.003676	7.683616
28	香皂	5	0.028249	0.003676	7.683616
29	意為	5	0.028249	0.003676	7.683616
30	怎會	5	0.028249	0.003676	7.683616

勝訴

	A	B	C	D	E
1		次數	頻率	敗-頻率	勝/敗率
2	所得額	16	0.059	0.000	--
3	機密	10	0.037	0.000	--
4	地理	9	0.033	0.000	--
5	賦與	8	0.029	0.000	--
6	重申	8	0.029	0.000	--
7	拉鍊	8	0.029	0.000	--
8	注重	7	0.026	0.000	--
9	米酒	7	0.026	0.000	--
10	至鉅	7	0.026	0.000	--
11	雙向	6	0.022	0.000	--
12	書包	6	0.022	0.000	--
13	得自	6	0.022	0.000	--
14	中段	6	0.022	0.000	--
15	旅行箱	6	0.022	0.000	--
16	食品類	5	0.018	0.000	--
17	人像	5	0.018	0.000	--
18	香煙	5	0.018	0.000	--
19	預先	15	0.055	0.006	9.761
20	抽象	10	0.037	0.006	6.507
21	如期	10	0.037	0.006	6.507
22	罐	10	0.037	0.006	6.507
23	判准	10	0.037	0.006	6.507
24	報稅	10	0.037	0.006	6.507
25	尚可	10	0.037	0.006	6.507
26	混為一談	10	0.037	0.006	6.507
27	飲品	9	0.033	0.006	5.857
28	原件	9	0.033	0.006	5.857
29	時日	9	0.033	0.006	5.857
30	市售	9	0.033	0.006	5.857

資料——判決內容處理

商品類別

18. 飲食類	26	19	0.731
19. 菸酒類	28	25	0.893
20. 車類	32	27	0.844
21. 科技產品類	32	12	0.375

24

- 案件詞頻

勝訴案件	次數	勝訴詞頻	敗訴詞頻	勝敗詞頻比
菸酒	12	0.027	0.009	3.000
菸品	8	0.018	0.004	4.000
機車	8	0.018	0.004	4.000
茶飲	6	0.013	0.004	3.250
敗訴案件				
充電器	4	0.009	0.004	2.250

- 分別勝敗訴案件計算詞頻後，結果發現大部分為**特定種類之商品**有相對高的出現頻率。飲食類+　菸酒類+　車類+　科技產品類 -

估計結果

- 加入斷詞
- 結果類似

	無商品分類		**加入商品分類**	
賠償金額				
刑附民	1.932	(1.297)	**2.553****	(0.984)
提到銷售量	1.039	(2.879)	1.200	(2.648)
故意	2.118+	(1.150)	**2.673****	(0.893)
過失	1.111	(1.737)	1.950	(1.339)
混淆誤認	-0.605	(1.005)	-0.253	(0.832)
著名商標	-2.522*	(1.058)	**-2.488***	(0.974)
民法	1.644*	(0.714)	1.638*	(0.661)
截距項	12.546***	(2.377)	12.698***	(2.523)
勝訴				
刑附民	0.511**	(0.190)	**0.469***	(0.193)
提到銷售量	0.077	(0.519)	0.086	(0.528)
故意	0.457*	(0.181)	**0.457***	(0.184)
過失	0.734*	(0.308)	**0.767***	(0.308)

◻ 估計結果

- 加入斷詞
- 結果類似
- Inverse Mills ratio 顯著

	Heckman_twostep		Heckman_斷詞	
混淆誤認	0.328*	(0.154)	**0.307+**	(0.158)
著名商標	0.105	(0.196)	0.100	(0.201)
民法	-0.027	(0.128)	-0.019	(0.131)
善意與合理使用	-0.905***	(0.243)	**-0.846*** **	(0.245)
斷詞 (飲食類)			0.298	(0.285)
斷詞 (菸酒類)			**1.037** **	(0.328)
斷詞 (車類)			**0.604***	(0.288)
斷詞 (科技產品類)			-0.324	(0.267)
截距項	0.074	(0.115)	-0.010	(0.120)
inverse Mills ratio	-4.877+	(2.927)	-2.943+	(1.593)
觀察值	449		449	

註：*表示統計上的顯著性：+為10%顯著水準，*為5%顯著水準，**為1%顯著水準，***為0.1%顯著水準。

◻ 小結

- 實證分析，對判決結果有影響力的審酌因素**並沒有那麼多**。
- 「刑附民之案件」、「混淆誤認」、「故意」和「過失」正向顯著影響勝敗率。
- **「善意及合理使用」**是顯著之排除變數，且只對勝敗訴造成影響。
- **特殊之商品**種類對於勝敗訴有顯著影響力。如菸酒、車類之產品遭受侵權時，更容易勝訴，但對賠償金額沒有影響。

◻ 利用非線性分群與判決文本能否讓勝訴預測更好?

- **Support-vector Machines** (支援向量機, SVM)
- **Word2Vec**
- 斷詞 + 勝敗詞頻倍率挑選 (倍率>2 或小於 0.5)

✓A0 重要變數→(Logit)
✓A1: A0重要變數→(SVM)
✓A2: A0重要變數→(決策樹)
✓B0: A0重要變數+特殊商品類型→(Logit)
✓B1: A0重要變數+特殊商品類型→(SVM)
✓B2: A0重要變數+特殊商品類型→(決策樹)

✓**C0,判決文本**法官心證(理由)部分 (Word2Vec, SVM)

◻ (A0) Logit vs. (A1) SVM支援向量機

- Logit模型預測：**準確率=0.6437**。

	模型預測 (大於平均值0.6058)		
實際資料	敗	勝	Total
敗	**126**	51	177
勝	109	**163**	272
Total	235	214	**449**

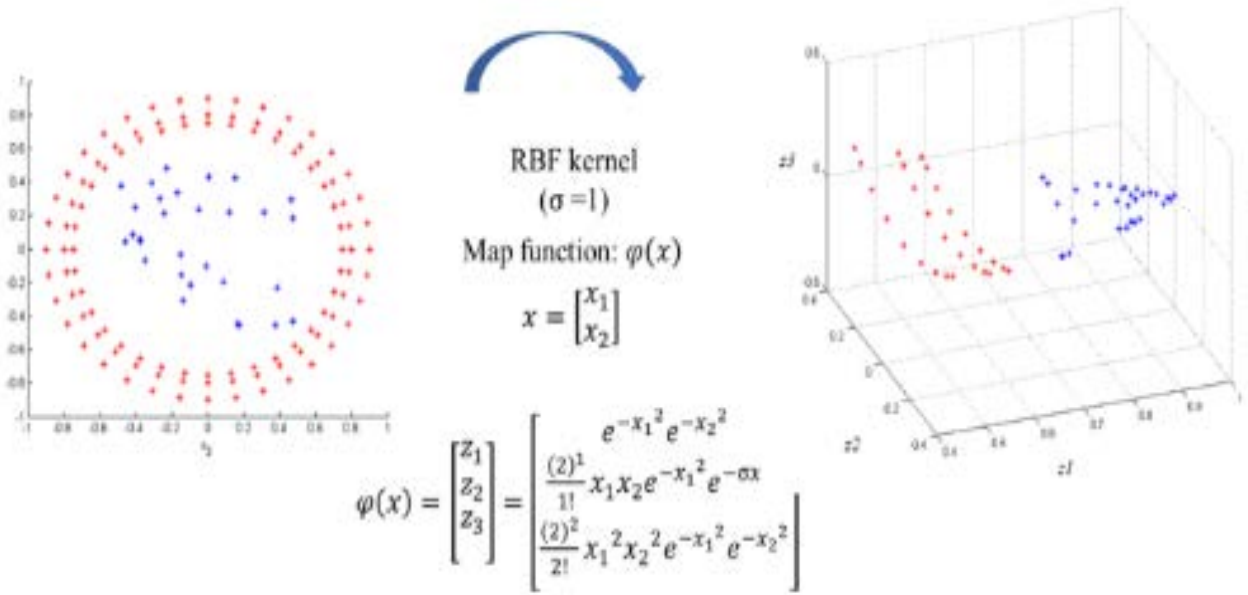
SVM預測：**準確率=0.6437**。(敗: 26/177, 勝: 263/272)
(Skip-gram, CBOW, linear, rbf, sigmoid)

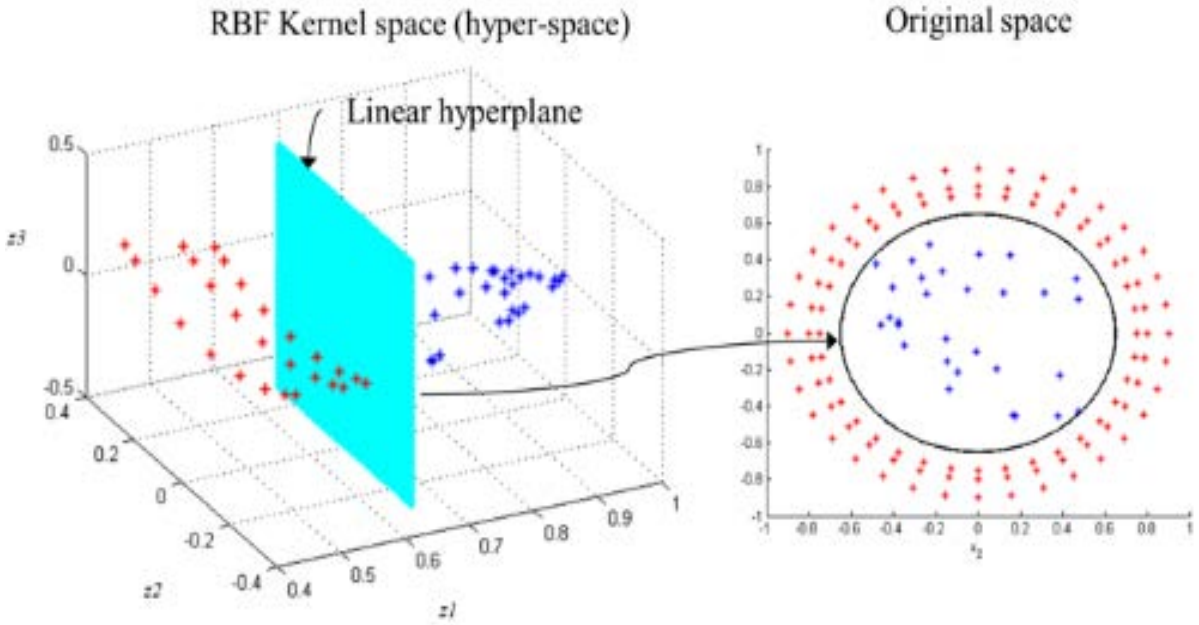
◻ A1: 重要變數 – SVM

SVM 核函數	維度	8
linear	勝訴	263/272
	敗訴	26/177
	正確率	**0.6437**
rbf	勝訴	263/272
	敗訴	26/177
	正確率	**0.6437**
sigmoid	勝訴	263/272
	敗訴	26/177
	正確率	**0.6437**

Rbf (Radial basis function kernel)

$$K(x_i, x_j) = exp(-\gamma||x_i - x_j||^2), \gamma > 0$$





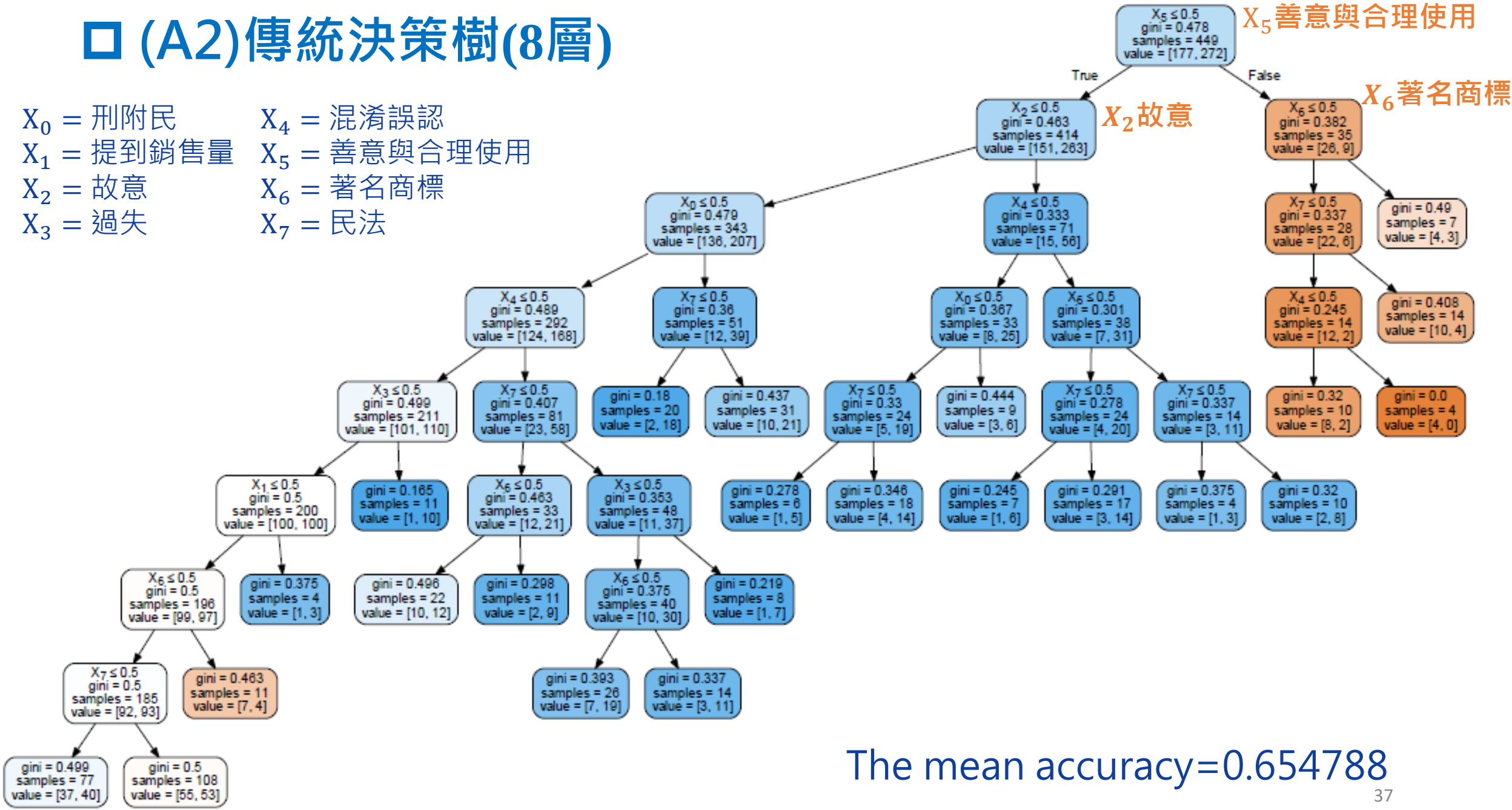
sigmoid

$$K(x_i, x_j) = tanh(\gamma x_i^T x_j + r), \gamma > 0, r < 0$$

資料來源： https://chih-sheng-huang821.medium.com/機器學習-kernel-函數-47c94095171

□ (A2)傳統決策樹(8層)

X_0 = 刑附民　　X_4 = 混淆誤認
X_1 = 提到銷售量　　X_5 = 善意與合理使用
X_2 = 故意　　X_6 = 著名商標
X_3 = 過失　　X_7 = 民法



The mean accuracy=0.654788

(B0)Logit vs.(B1)SVM支援向量機—重要變數+特殊商品

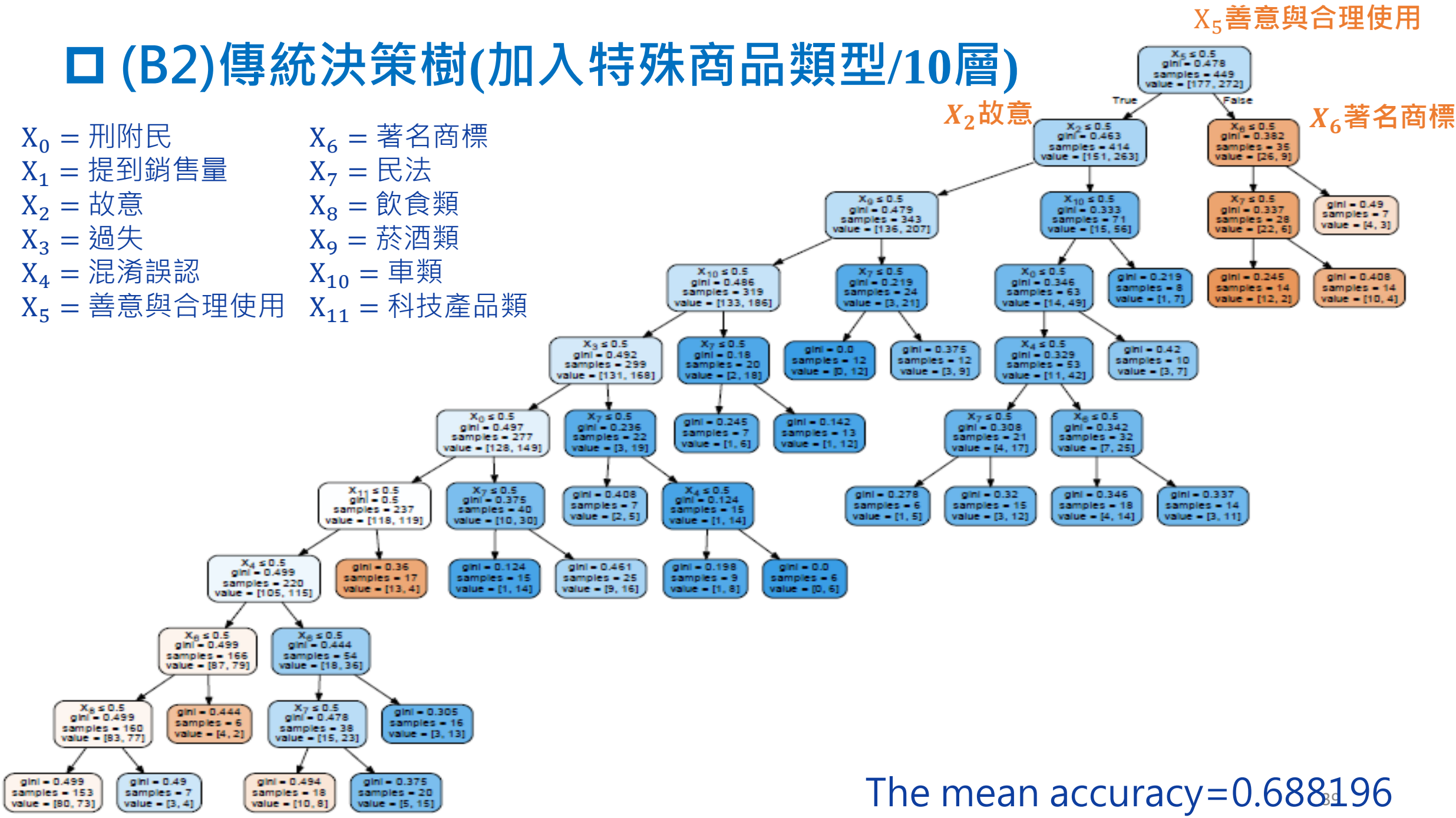
- Logit模型預測：**準確率=0.6793**。

	模型預測 (大於平均值0.6058)		
實際資料	敗	勝	Total
敗	**126**	51	177
勝	93	**179**	272
Total	219	230	**449**

- SVM預測：**準確率=0.6392~0.6437 (嘗試不同商品組合, 和不加時類似)**

◻ (B2)傳統決策樹(加入特殊商品類型/10層)

X_0 = 刑附民
X_1 = 提到銷售量
X_2 = 故意
X_3 = 過失
X_4 = 混淆誤認
X_5 = 善意與合理使用
X_6 = 著名商標
X_7 = 民法
X_8 = 飲食類
X_9 = 菸酒類
X_{10} = 車類
X_{11} = 科技產品類



The mean accuracy=0.688196

◻ 法官心證之理由部分－Word2Vec(Skip-gram)+SVM

SVM核函數	維度	25	50	75	100
linear	勝訴	216/272	207/272	209/272	210/272
	敗訴	84/177	89/177	84/177	98/177
	正確率	0.6682	0.6592	0.6526	**0.6860**
rbf	勝訴	212/272	214/272	213/272	217/272
	敗訴	97/177	90/177	83/177	90/177
	正確率	**0.6882**	0.6771	0.6592	**0.6837**
sigmoid	勝訴	215/272	212/272	217/272	229/272
	敗訴	82/177	83/177	85/177	57/177
	正確率	0.6615	0.6570	0.6726	0.6370

◻ (C0)法官心證(理由)部分－Word2Vec (CBOW) + SVM

SVM核函數	維度	25	50	75	100
linear	勝訴	218/272	216/272	234/272	230/272
	敗訴	79/177	85/177	77/177	84/177
	正確率	0.6615	0.6704	**0.6927**	**0.6993**
rbf	勝訴	236/272	225/272	249/272	225/272
	敗訴	66/177	75/177	67/177	86/177
	正確率	0.6726	0.6682	**0.7038**	**0.6927**
sigmoid	勝訴	225/272	223/272	234/272	230/272
	敗訴	80/177	69/177	77/177	84/177
	正確率	**0.6793**	0.6503	**0.6927**	**0.6993**

結果彙整

模型	全猜勝	重要變數 Logit & SVM (決策樹)	重要變數+特殊商品類型 Logit (SVM/決策樹)	(C)法官心證(理由)部分 – Word2Vec + SVM(CBOW)
正確率	0.6058	0.6437 (0.6548)	0.6793 (0.6392~0.6437 /0.6882)	0.7038
				V

結論

- 計量分析中，部分學術界建議的審酌因素因在統計上不具顯著性。
  - 「刑附民之案件」、「故意」、以及「涉及民法」皆會對賠償金額造成顯著正向影響，但「著名商標」對賠償金額影響為負向。
  - 「混淆誤認」正向顯著影響勝敗率，但對於賠償金額的影響不顯著。
  - 「提到故意或過失」、「善意及合理使用」是顯著之排除變數，只對勝敗訴造成影響。
- 大數據可以協助找出其他**非法學界歸納出**之影響因素: 特殊商品類型。
- **文字探勘**(斷詞+Word2Vec+SVM) 有機會額外增加正確率的判斷。
- 使用不同演算法，提高預測正確率？

◻ 利用XGBoost與LSTM能否再提升勝訴預測結果?

- **訓練集：75% (n=336)；測試集：25% (n=113) & 10次平均**

✓C1. 判決法官心證(理由)部分 (Word2Vec)—**XGBoost** (eXtreme Gradient Boosting)
✓**C2**. (Word2Vec+ A0重要變數+特殊商品)—**XGBoost**

✓D1. C1篩選變數+A0重要變數+特殊商品—**SVM**
✓**D2.** C1篩選變數+A0重要變數+特殊商品— **XGBoost**

✓E1. C2做變數篩選—**SVM**
✓E2. C2做變數篩選—**XGBoost**

- **LSTM(長短期記憶模型遞歸神經網路) (10次平均)**
  - 結合C1+C2篩選出之變數，以時間排序各分判決，前一筆預測下一筆

◻ XGBoost

- 學習目標函式

- Model: assuming we have K trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}$$

- Objective

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Training loss　　Complexity of the Trees

- 目標函式(保留了泰勒展開的二次項)

- Objective, with constants removed

$$\sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

  - where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}), \quad h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$

資料來源：https://codertw.com/程式語言/635146/

◻ XGBoost：Boosting

- Gradient Boosting 學習方式：每次預測時新加入一個學習函數 (參數)
- 再讓 $f_t(x_i)$ 對目標函數做優化

$$
\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\cdots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}
$$

新加入的函數

第t次預測

保留先前加入過的所有學習函數

資料來源： https://medium.com/@gary1346aa/機械學習-gradient-boosting-1-272e8d7b17e1

D2. XGBoost： C1【Word2Vec—XGBoost 】+A0重要變數+特殊商品

- 訓練集：75% (n=336)；測試集：25% (n=113)

	原w2v分類	25	50	75	100
Skip-gram	篩選後變數量	31	34	23	60
	勝訴正確率	0.7592	0.7351	0.7702	0.7736
	敗訴正確率	0.6591	0.7081	0.7306	0.6850
	整體正確率	0.7239	0.7257	**0.7566**	0.7407
CBOW	篩選後變數量	23	36	19	16
	勝訴正確率	0.7658	0.7564	0.7885	0.7397
	敗訴正確率	0.7563	0.6848	0.6990	0.6605
	整體正確率	**0.7628**	0.7319	**0.7566**	0.7133

◻ E1. SVM：C2【Word2Vec(Skip-gram)+A0重要變數+特殊商品—XGBoost】,

	原w2v分類	25	50	75	100
核函數	篩選後變數量	15	12	11	80
linear	勝訴正確率	0.7081	0.7320	0.7313	0.7583
	敗訴正確率	0.7551	0.6263	0.7301	0.7471
	整體正確率	0.7204	0.6965	0.7310	**0.7549**
rbf	勝訴正確率	0.7705	0.7467	0.7461	0.7611
	敗訴正確率	0.6943	0.6649	0.7224	0.6935
	整體正確率	0.7469	0.7195	0.7398	0.7381
sigmoid	勝訴正確率	0.7377	0.7026	0.7276	0.7449
	敗訴正確率	0.6944	0.6629	0.7730	0.6943
	整體正確率	0.7248	0.6903	0.7398	0.7292

◻ 結果彙整

模型	(C1)XGBoost：大數據探勘下法官心證(理由)部分		(C2)XGBoost：大數據探勘下法官心證(理由)部分 +重要變數+特殊商品類型	
正確率	0.7195		0.7416	
模型	(D1)C1+重要變數+特殊商品類型 SVM	(D2)C1+重要變數+特殊商品類型 XGBoost	(E1)C2-SVM	(E2) C2-XGBoost
正確率	0.7575	**0.7628**	0.7593	0.7611

結果

- (D2) [C1+重要變數+特殊商品類型] XGBoost 預測正確率：**0.7628**
- 變數重要性排序

	1	2	3	4	5	6	7	8	9	10	11
Feature	**w16**	**w8**	**w19**	**w9**	過失	w10	w20	w13	故意	w6	w21
Weight	**0.5828**	**0.2514**	**0.2072**	**0.1856**	0.081	0.0762	0.0594	0.0564	0.0433	0.041	0.0387
	12	13	14	15	16	17	18	19	20	21	22
Feature	菸酒類商品	混淆誤認	w22	w1	w23	w5	w12	w3	車類商品	善意與合理使用	w25
Weight	0.0356	0.0348	0.0298	0.0286	0.0253	0.0246	0.0239	0.0222	0.0187	0.0123	0.0108
	23	24	25	26	27	28	29				
Feature	w18	w17	銷售量	w24	著名商標	科技產品類商品	民法				
Weight	0.0087	0.0031	0.0023	0.0022	0.002	0.0013	0.0006				

結果

- 變數重要性排序 (以向量排序，且屬於重要詞頻)

	最大,勝/敗100內前4,勝/敗100內次數>10 前2	敗/勝100內次數>10 前2 ,敗/勝100內前4, 最小
w16	包(最大), 公事包, 稅務, 香菸, 抽象, 不及, 稅務,	承辦, 新舊, 開箱, 場地, 同仁, 染 (最小)
w8	英語(最大), 不及, 如期, 息息相關, 錄影	承辦, 移轉給, 通關, 淘, 染, 上訴人(最小)
w19	鎖(最大), 塑膠, 香菸, 公事包, 電話號碼, 不詳	借款, 計劃, 加盟(最小)
w9	登載(最大), 修理, 錄影, 不詳, 公事包	現存, 臚列, 手寫, 為期, 上訴人(最小)
w10	不詳(最大), 教師, 香菸, 修理	承辦, 新舊, 為期, 現存, 禾, 傳真機(最小)
w20	費用(最大), 借款, 計劃, 反光, 報稅, 稅捐	承辦, 保養品, 緩, 袋包, 染(最小)

◻ Kmeans 分群 (24)　10次以上　勝/敗100內　敗/勝 100內

勝/敗100內	10次以上	敗/勝 100內	10次以上
擔保金	2	預先	1	新舊	13
稅捐	3	報稅	10		
稅務	4	列舉	51	承辦	37
為最	6	語文	74		
罐	7	舉出	89	保養品	97
抽象	9	尚可	8		
判准	11				
混為一談	12	貶值	87	W16 W8 W19 W9 W10 W20	
如期	13	塑膠	91	W13 W6 W21 W18 W17 W24	
估算	53				
不及	54	不詳	52		
年籍	88	修理	5		
屆	90				

討論

1.審酌因素
法學領域的整理
領域外的因素
標定工程
(工人,人工智慧)
詞彙, 詞彙組?

2. 斷詞及語意空間
個人電腦?
依案件類別?
詞庫的建立
與分享

3. 事實的認定
法官認定的事實vs.
當事人覺得的事實
判決書中反駁
與認定

4. 新的人工智慧方法, 組合, 應用

謝謝聆聽