【台灣法實證研究資料庫活動】

機器學習在法律上的應用工作坊

文字紀錄：江明峰、鍾予晴

時間：2015 年 1 月 28 日 13:00-17:00
地點：法律學院霖澤館七樓 1710 第一會議室
主持：陳昭如（國立台灣大學法律學院教授）
指導：林守德（國立台灣大學資訊工程學系副教授）
講者：駱柏蓁（Intel-台大創新研究中心 研究助理）
陳煥元（台大資工系）

這一個工作坊的主旨是探討：電腦是否能夠模擬法律人的思考，而法律人的思考是否能被電腦模仿。透過電腦的模擬，可以歸納大量的資訊，也可以嘗試對未來的事件作出可能的判斷。

一、人工智慧簡介

在談電腦模擬法律人的思考之前，要先了解的背景知識是「人工智慧」（Artificial Intelligence，簡稱 AI）。人工智慧就是：希望電腦程式具備像人一樣的智慧，可以理解人類的語言、回答問題、從經驗中學習並發現新的事物定理、自己解決問題等。

人工智慧的領域分成兩支：強人工智慧與弱人工智慧。前者希望電腦可以與人的思維模式接近或相同，弱人工智慧則希望電腦可以「有智慧的」作出一些表現或成果。

強人工智慧是希望電腦有智慧，這麼一來，就要了解人類自己是怎麼思考的，並讓電腦盡可能模擬，但這是比較難發展的方向。我們可以這樣比喻強人工智慧：要求電腦跟人腦一樣，就像是要求人類像鳥一樣地飛，這相當困難。相對地，弱人工智慧是希望人類會飛，但不在乎人們怎麼飛。而機器學習是弱人工智慧的方法中最有效的一種。

機器學習（Machine Learning）是一種讓電腦自動從資料中學習的技術，其中最重要的兩個要素就是歸納與推論。所謂歸納方法，舉例而言，我們今天看了很多案件事實，並依據所判的罪刑去分類和分析判決，我們發現通常有「使用武器」、「死亡」、「逃逸」等情況的事實會是殺人罪。而事實中包含「使用武器」、「受傷」等情況的會是傷害罪，透過這樣的方式我們可以建立知識。推論方法是運用

既有的前提或背景知識，例如，前面寫說「使用武器」、「死亡」、「逃逸」等情況會是殺人罪，那「沒用武器」、「死亡」的情況是否也能推論是殺人罪？恐怕不行，因為背景知識無法支持這個推論。機器學習基本上就是讓電腦也能夠使用歸納與推論。電腦並不是真的去理解死亡、逃逸的意思是甚麼，不過可以用統計的方式去歸納推論它們。

所謂「機器學習研究」是研究「如何讓電腦從大量資料中學習知識」的一門科學。例如，我們可以從大量(x,y)的資料中來預測 y=f(x)的 f(x)是甚麼。

機器學習的種類，今天會著重在規則式學習（Rule Based Learning），例如有一個人的行為表現是「吃多、喝多、尿多」，那照規則來看，他就可能是糖尿病。規則式學習的優點是容易了解與建立，缺點是需要仰賴該領域的專家、規則不容易完備、建立系統曠日廢時（因為該領域的專家不一定懂電腦），並且，不同的人，他的判斷規則也不同，因而難以統一這些規則。

機器學習可以用來做分類與分群、預測、關聯、解釋，在法律上的應用有判決書自動標記、犯罪類型判別、犯罪量刑預測。它利用了電腦善於計算的長處，從大量資料中找出規則、關聯，並做出判斷和預測，可以記憶大量資料、分析重要因子，並避免偏見。

二、機器學習入門

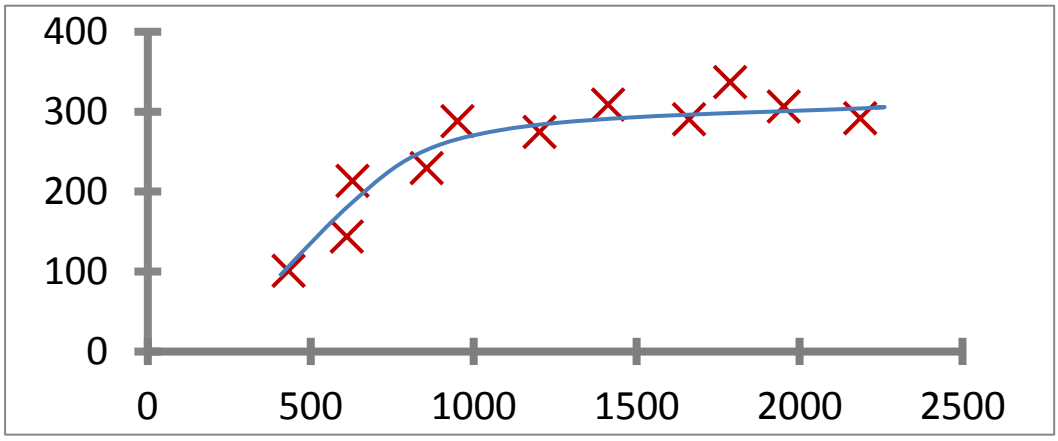
機器學習是提供大量的資料給電腦，讓電腦學習之後得以做出預測和判斷。機器學習的應用非常廣泛，例如臉書上的人臉辨識功能、自動駕駛、手寫辨識功能等等，都是讓電腦接收大量的資料並學習判讀。

機器學習可以分為兩大類：監督式學習（supervised learning）及非監督式學習（unsupervised learning）。監督式學習是給電腦一堆「有標籤的資料」（labeled data）並給電腦明確的指示，讓電腦按照指示來分析資料。非監督式學習則是給電腦許多「未標籤的資料」（unlabeled data），並讓電腦依照簡單的指示自動進行分類。

這次要介紹的是監督式學習方法。監督式學習法運用的是「有標籤的資料」（labeled data），可以用有這些有標籤的資料作迴歸（regression）和決策樹（decision tree）等方面的分類和分析。

迴歸（regression）

假設我們打算探討房子大小和房價的關係，因此蒐集了許多房子大小和房子的資料，將這些資料表示成(x, y)的形式，其中 X 表示房子大小，y 表示房價。許多的(x, y)畫在 xy 軸上，形成許多資料點。而迴歸就是找出一條線，而所有資料點與那條線的距離平方和呈最小值。而這一條線就能代表房子大小和房價的關係。



*圖中的藍色線為迴歸線，用以表示房子大小和房價的關係。

決策樹（decision tree）

決策樹所運用的資料，變數只有「是」或「否」兩種情形（迴歸分析中，資料則是連續性的數值）。決策樹如何建立？我們必須先找出一個變因，將資料依據該變因分成兩組，終極目標是同組內資料變異最小化，組間的變異最大化，以達到分類效果。我們可以讓電腦學習找出變因來進行分類。

為了要得知決策樹是否有成功達到分類的效果，我們需要量化的指標。Entropy 值就是評估決策樹分類效果如何的指標，值越小表示分類效果越好。Entropy 值的計算公式如下：

$$E = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

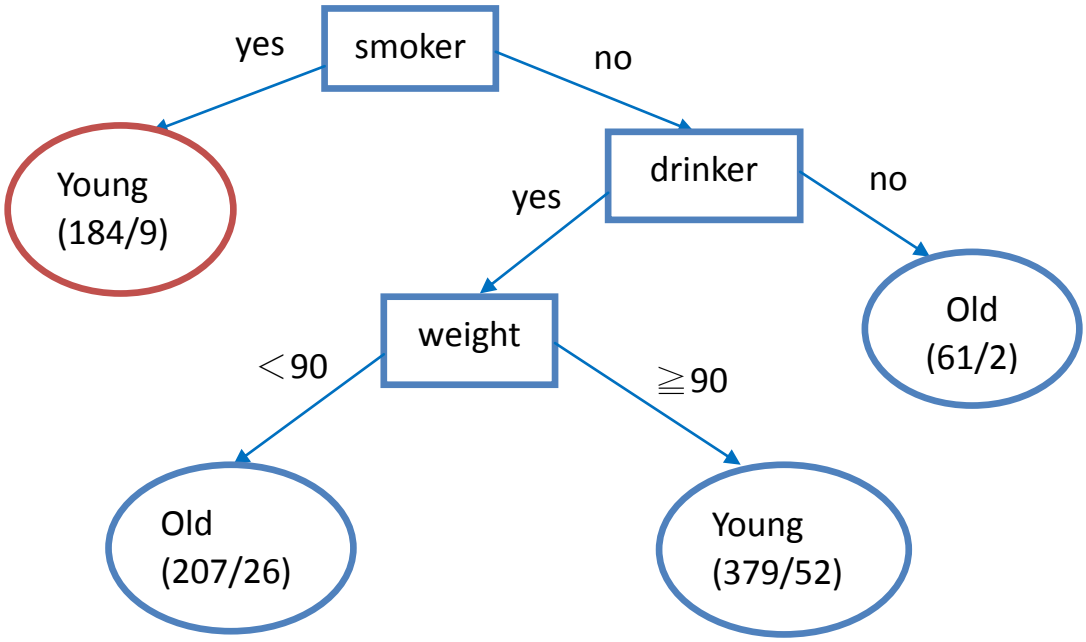
決策樹的建立方式，舉例說明：

假設我們有 831 筆關於吸菸、喝酒、年齡和體重的資料，部分資料如下表：

drink	smoker	weight	age
Yes	Yes	120	44
No	No	70	96
Yes	No	72	88
Yes	Yes	55	52
No	Yes	94	56
:	:	:	:

*age＜70 為年輕（Young），age≧70 為年老（old）。

接著我們找出一一找出適當的變因，來一層層的分類資料。這些變因依序是「是否為吸菸者（smoker）」、「是否有飲酒習慣（drinker）」、「體重是否大於 90 公斤」，如下圖的方框。分類出的資料，我們再依照資料呈現的年齡概況，標示為年輕（young）或年老（old），如下圖的圓框。



但不是全部的資料都可以正確地被標示為年輕或年老。例如左上紅色圓框中，共有 184 筆資料被分類在這個區塊，其中有 175 筆資料的年齡是「年輕」，但有 9 筆資料其實年齡是「年老」。由於這個圓框中的資料絕大多數是屬於「年輕」，所以該圓框被標為 young。

基本上，資料不太可能完全正確的被分類，因此只能追求組內的差異盡量減少，讓我們做預測時可以更加準確。

評估電腦的學習成果

當電腦根據大量的資料建立出迴歸線或決策樹後，可能會有「模型太符合現有資料」（overfitting）和「模型不夠符合資料」（underfitting）等情況。「模型太符合現有資料」（overfitting）意味著：未來有新的資料進來時，迴歸線或決策樹的預測效果會變差。「模型不夠符合資料」（underfitting）則是指，迴歸線或決策樹設得太簡單了，以致無法預測。實務上常發生「模型太符合現有資料」的情況，導致未來做分類的效果並不佳。為了避免 overfitting 的情況發生，可以將某份大筆資料切成 K 份，並分成訓練電腦用的資料（training set）和測試模型用的資料（test set），K 值越大越好。

三、實作過程簡介

1. 找出關鍵詞：案件事實中，哪些事實是影響判決的關鍵？

   A. 使用資料：8 篇強盜罪的判決。

   B. 斷詞

   所謂斷詞就是把一段文字變成一個一個詞。斷詞後，需要將這些語詞作詞性標記，以便接下來的資料分析。而其可能的應用有，自動分類裁判書、自動標記影響判決之關鍵詞彙、標註事實與理由關鍵詞彙的配對、自動找尋理由的法源依據。

   這次實作的資料是判決書，判決書格式有主文、事實、理由，事實是比較客觀的部分，所以可以用來作分析。在「斷詞」這個步驟中，要將事實裡重要關鍵字標記起來。目前有些學術機構有提供斷詞系統供大家使用，例如中研院斷詞系統。這個步驟就是將判決文字放進中研院斷詞系統中，把文句才成語詞。

   中研院斷詞系統的操作結果如圖所示：

```
主文(Na) 蕭宏傑(Nb) 意圖(VF) 為(P) 自己(Nh) 不法(A) 之(DE) 所有(Neqa) ，(COMMACATEGORY)
------------------------------------------------------------------------------------------
結夥(D) 三(Neu) 人(Na) 以上(Ng) ，(COMMACATEGORY)
------------------------------------------------------------------------------------------
攜帶(VC) 兇器(Na) ，(COMMACATEGORY)
------------------------------------------------------------------------------------------
以(P) 強暴(VC) 至(Caa) 使(VL) 不能(D) 抗拒(VC) ，(COMMACATEGORY)
------------------------------------------------------------------------------------------
而(Cbb) 取(VC) 他人(Nh) 之(DE) 物(Na) ，(COMMACATEGORY)
------------------------------------------------------------------------------------------
處(Nc) 有期徒刑(Na) 捌年(Nd) 陸月(Nb) ；(SEMICOLONCATEGORY)
------------------------------------------------------------------------------------------
又(D) 共同(A) 傷害(Na) 人(Na) 之(DE) 身體(Na) ，(COMMACATEGORY)
------------------------------------------------------------------------------------------
處(Nc) 有期徒刑(Na) 肆月(VC) 。(PERIODCATEGORY)
------------------------------------------------------------------------------------------
應(D) 執行(VC) 有期徒刑(Na) 捌年(Nd) 捌月(Nd)
------------------------------------------------------------------------------------------
```

   C. 關鍵詞自動標記

   這個步驟要運用的原理是條件隨機場（CRF），就是把能看到的詞全部納入考慮，將機率算出來。要使用 Mac OS 系統的終端機（Terminal）或 Windows 系統的命令提示字元（command line, cmd）來操作。

   如前面的機器學習簡介所說明的，我們可以先用部分資料去作訓練，再用部分資料去作測試，讓電腦跑出合適的模型。

   訓練的指令如下：

   crf_learn template train [model] (mac)

   crf_learn.exe template train [model] (windows)

   [model]為自訂模型檔名

   測試的指令如下：

   crf_test –m [model] test > [output] (mac)

   crf_test.exe –m [model] test > [output] (windows)

[model]為已照前一步驟訓練完畢的模型檔名

2. 判決文分類：建立決策樹，找出分類強盜罪和恐嚇取財罪的變因。

A. 使用資料：

判決書事實段落的詞頻分析。

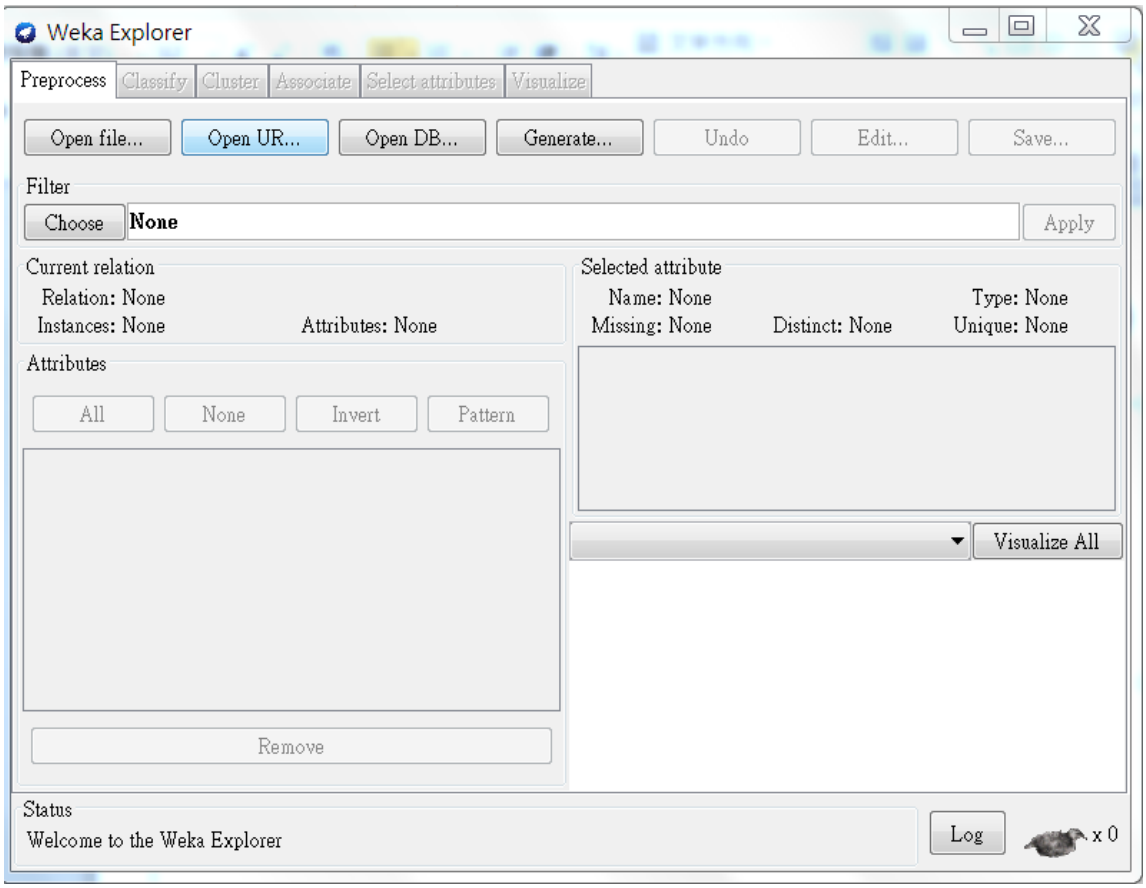
全部資料：28 篇強盜判決書，20 篇恐嚇取財判決書。

用以訓練機器的資料（train）：20 篇強盜判決書，12 篇恐嚇取財判決書。

用以測試模型的資料(test)：8 篇強盜判決書，8 篇恐嚇取財判決書。

B. 決策樹分析

這個步驟要使用的軟體是 weka，完成了安裝程序後，要先進入預處理（preprocess）的介面，將資料作轉檔的處理，把 csv 格式轉成 arff 格式，介面如下圖：



處理好資料後，可以進入分類（classify）的介面，將資料進行處理，建立出決策樹，介面如下：