附錄

附錄一　中研院中文詞性標記集對照表

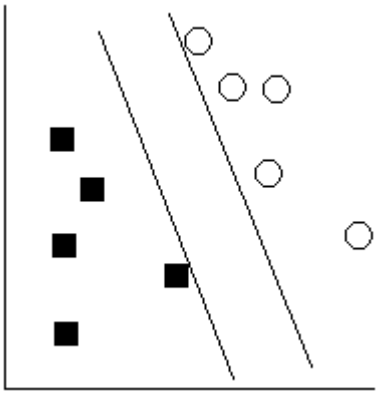
簡化詞類	代表意義	精簡詞類	簡化詞類	代表意義	精簡詞類	簡化詞類	代表意義	精簡詞類
A	非謂形容詞	A	Nb	專有名稱	N	VB	動作類及物動詞	Vi
Caa	對等連接詞	C	Nc	地方詞	N	VC	動作及物動詞	Vt
Cab	連接詞，如：等等	POST	Ncd	位置詞	N	VCL	動作接地方賓語動詞	Vt
Cba	連接詞，如：的話	POST	Nd	時間詞	N	VD	雙賓動詞	Vt
Cbb	關聯連接詞	C	Nep	指代定詞	DET	VE	動作句賓動詞	Vt
D	副詞	ADV	Neqa	數量定詞	DET	VF	動作謂賓動詞	Vt
DE	的, 之, 得, 地	T	Neqb	後置數量定詞	POST	VG	分類動詞	Vt
Da	數量副詞	ADV	Nes	數量副詞	DET	VH	狀態不及物動詞	Vi
Dfa	動詞前程度副詞	ADV	Neu	數詞定詞	DET	VHC	狀態使動動詞	Vt
Dfb	動詞後程度副詞	ADV	Nf	量詞	M	VI	狀態類及物動詞	Vi
Di	時態標	ASP	Ng	後置	POST	VJ	狀態	Vt

	記			詞			及物動詞	
Dk	句副詞	ADV	Nh	代名詞	N	VK	狀態句賓動詞	Vt
FW	外文標記	FW	SHI	外文標記	Vt	VL	狀態謂賓動詞	Vt
I	感嘆詞	T	T	語助詞	T	V_2	有	Vt
NAV	名謂詞	NAV	VA	動作不及物動詞	Vi			
Na	的, 之, 得, 地	N	VAC	動作使動動詞	Vi			

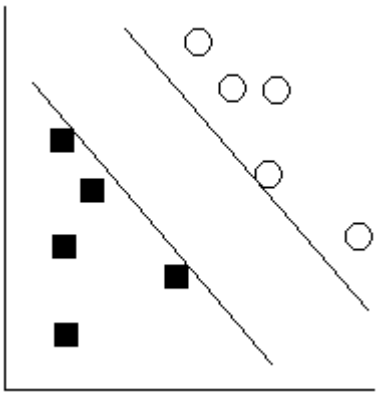
附錄二 支持向量機 Support Vector Machine (SVM) 簡介

SVM 是較新的 machine learning 技術 (Boser, Guyon, and Vapnik (1992), Corts and Vapnik (1995)) 它使用一些策略來最大化具有不同特徵的資料中間的界限, 並針對未知資料的特徵來判斷它屬於哪個類別。SVM 已在文件分類 (Joachims (1998) Taira and Haruno (1999)) 以及名詞組標示 (Kudo and Matsumoto (2000, 20001))取得超越其它作法的準確性, 而近幾年應用在自然語言處理的各個議題的研究更是方興未艾，如未知詞辨識 (unknown word guessing) (Nakagawa, Kudo, and Matsumoto (2001)) 詞性標注(part of speech tagging)Nakagawa, Kudo, and Matsumoto (2002)， Giménez J esús and Márquez Lluís (2004))句法依存關係辨識(dependency analysis)(Kudo and Matsumoto (2000))詞義辨別與標注(word sense disambiguation and sense tagging) (Cabezas, Resnik, and Stevens (2001))語意剖析(semantic parsing) (Pradhan et al. (2004) Sun and Jurafsky (2004))等都取得不錯的成果。

SVM 是一個分類用的 machine。請參照圖 (一，二),



圖一



圖二

SVM 找出兩種資料 (黑色方形與白色圓形) 中間的界限, 圖一，圖二顯示出可能的兩種分割方式, 顯然的, 後者的切割方式是較佳的 (兩種資料的界線為兩平行線之中線), 而 SVM 以滿足下面條件

$$\min\Phi(\omega)= (1/2)|\omega|^2$$

找出最佳平面（即在線性可分的情況下，可視為解二次規畫的問題），而此可由拉格朗日乘子法（Lagrange multiplier）求解。

由於很多的問題常常並不是線性可分的（如我們的詞組切割），這個時候 SVM 在比現有資料更高的向量空間 H 使用線性分類函數 $\Phi: R^d \rightarrow H$ 將 x 對應到高維空間，便可

在此以不破壞資料特徵亦不增加複雜度的方式對其進行分類.在轉換的過程中，我們會使用一 kernel function: $K(x_i,x_j) = \Phi(x_i)^T\Phi(x_j)$ 來實現非線性變換後的線性分類，而使用不同的 kernel function 對不同的資料會有不同的效果.以下為一個簡單的SVM 運作方式

給定一個訓練的資料集合:

(x_i, y_i) { i= 1, 2, ..., l; x_i 屬於 R^n; y_i 屬於 { 1, -1 }}

其中 l 為訓練之資料數，x_i 為一個 n 維向量，y_i 則是其類別（分為正類別 1 與負類別 -1）SVM 找到正類別與負類別中之最大的界限，即解決下面的最佳化問題的解答

$$\min_{w,b,e}(1/2)w^Tw + C\sum_{i=1}^{l}e_i \text{ 使得}$$
$$y_i(w^T\Phi(x_i) + b) >= 1- e_i,\ e_i >= 0$$

x_i 經由 Φ 函數被對應到一個更高維的向量空間 H之後 SVM 於此找到不同類別之間最大的界限; $K(x_i,x_j)$ 為 Kernel function.

附錄三 Bayesian Classification 簡介

以下簡述 Bayesian Classification。假設我們現在要對一個目標詞做詞義辨認，該目標詞的詞義有 k 個，依序是$s_1, s_2, \ldots, s_k$，則目標就是要找出一個s'，使得$P(s'|c)$為最大，c 是目標詞所含有的某種特徵。根據貝式定理，可以得到如下的等式：

$$P(s_k|c) = \frac{P(c|s_k)}{P(c)}P(s_k)$$

因此

$$s' = \arg\max_{s_k} P(s_k|c)$$

$$= \arg\max_{s_k} \frac{P(c|s_k)}{P(c)}P(s_k)$$

$$= \arg\max_{s_k} P(c|s_k)\,P(s_k)$$

$$= \arg\max_{s_k} [\log P(c|s_k) + \log P(s_k)]$$