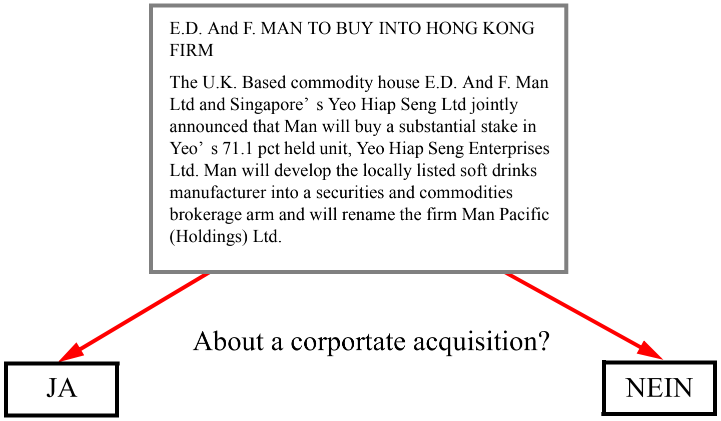
Linear Classifiers and Perceptron

CS678 Advanced Topics in Machine Learning
Thorsten Joachims
Spring 2003

Outline:

- Linear classifiers
- Example: text classification
- Perceptron learning algorithm
- Mistake bound for Perceptron
- Separation margin
- Dual representation

Text Classification

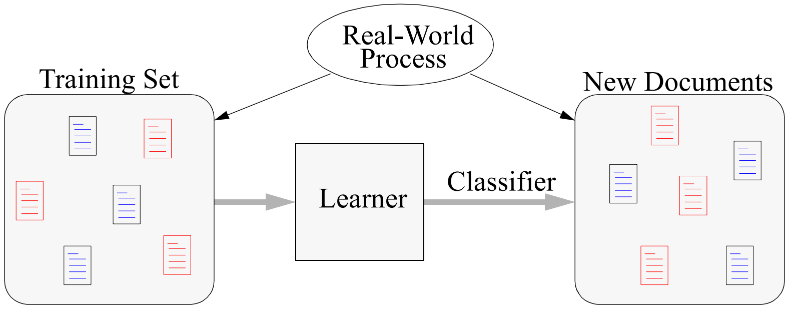


E.D. And F. MAN TO BUY INTO HONG KONG FIRM

The U.K. Based commodity house E.D. And F. Man Ltd and Singapore' s Yeo Hiap Seng Ltd jointly announced that Man will buy a substantial stake in Yeo' s 71.1 pct held unit, Yeo Hiap Seng Enterprises Ltd. Man will develop the locally listed soft drinks manufacturer into a securities and commodities brokerage arm and will rename the firm Man Pacific (Holdings) Ltd.

About a corportate acquisition?

Learning Text Classifiers



Goal:

- Learner uses training set to find classifier with low prediction error.

Generative vs. Discriminative Training

Process:

- Generator: Generates descriptions $\vec{x}$ according to distribution $P(\vec{x})$.
- Teacher: Assigns a value y to each description $\vec{x}$ based on $P(y|\vec{x})$.

=> Training examples $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \sim P(\vec{x}, y)$ $\quad \vec{x}_i \in \Re^N$ $\quad y_i \in \{1, -1\}$

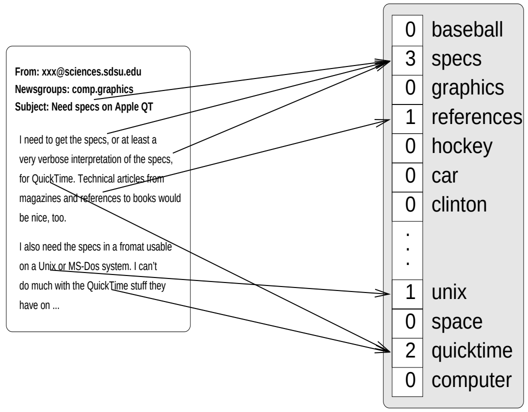
Generative Training

- make assumptions about the parametric form of $P(\vec{x}, y)$.
- estimate the parameters of $P(\vec{x}, y)$ from the training data
- derive optimal classifier using Bayes' rule
- example: naive Bayes

Discriminative Training

- make assumptions about the set H of classifiers
- estimate error of classifiers in H from the training data
- select classifier with lowest error rate
- example: SVM, decision tree

Representing Text as Attribute Vectors



From: xxx@sciences.sdsu.edu
Newsgroups: comp.graphics
Subject: Need specs on Apple QT

I need to get the specs, or at least a very verbose interpretation of the specs, for QuickTime. Technical articles from magazines and references to books would be nice, too.

I also need the specs in a fromat usable on a Unix or MS-Dos system. I can't do much with the QuickTime stuff they have on ...

0	baseball
3	specs
0	graphics
1	references
0	hockey
0	car
0	clinton
. . .	
1	unix
0	space
2	quicktime
0	computer

Attributes: Words (Word-Stems)

Values: Occurrence-Frequencies

==> The ordering of words is ignored!

Linear Classifiers (Example)

Text Classification: Physics (+1) versus Receipes (-1)

ID	nuclear (x_1)	atom (x_2)	salt (x_3)	pepper (x_4)	water (x_5)	heat (x_6)	and (x_7)	y
D1	1	2	0	0	2	0	2	+1
D2	0	0	0	3	0	1	1	-1
D3	0	2	1	0	0	0	3	+1
D4	0	0	1	1	1	1	1	-1

w,b	2	3	-1	-3	-1	-1	0	b=1

$$\text{D1: } \sum_{i=1}^{7} \vec{w}_i \vec{x}_i + b = [2\cdot1+3\cdot2+(-1)\cdot0+(-3)\cdot0+(-1)\cdot2+(-1)\cdot0+0\cdot2]+1$$

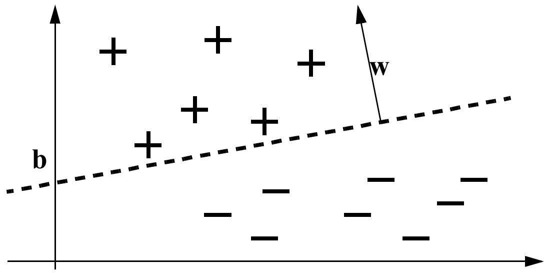
$$\text{D2: } \sum_{i=1}^{7} \vec{w}_i \vec{x}_i + b = [2\cdot0+3\cdot0+(-1)\cdot0+(-3)\cdot3+(-1)\cdot0+(-1)\cdot1+0\cdot1]+1$$

Linear Classifiers

Rules of the Form: weight vector $\vec{w}$, threshold b

$$h(\vec{x}) = sign\left[\sum_{i=1}^{N} \vec{w}_i \vec{x}_i + b\right] = \begin{cases} 1 & if \sum_{i=1}^{N} \vec{w}_i \vec{x}_i + b > 0 \\ -1 & else \end{cases}$$

Geometric Interpretation (Hyperplane):



Perceptron (Rosenblatt)

Input: $S = {\circledR}(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n) \lfloor \quad \vec{x}_i \in \Re^N \quad y_i \in \{1, -1\}$ (linear separable)

```
• $w_0 \leftarrow 0; b_0 \leftarrow 0; k \leftarrow 0$
• $R = max_i \|\vec{x}_i\|$
• repeat
    • for i=1 to n
        • if $y_i(\vec{w}_k \cdot \vec{x}_i + b_k) \leq 0$
            • $\vec{w}_{k+1} \leftarrow \vec{w}_k + \eta y_i \vec{x}_i$
            • $b_{k+1} \leftarrow b_k + \eta y_i R^2$
            • $k \leftarrow k = 1$
        • endif
    • endfor
• until no mistakes made in the for loop
• return $(\vec{w}_k, b_k)$
```

Analysis of Perceptron

Definition (Margin of an Example): The margin of an example $(\vec{x}_i, y_i)$ with respect to the hyperplane $(\vec{w}, b)$ is

$$\delta_i = y_i(w \cdot x_i + b)$$

Definition (Margin of an Example): The margin of a training set $S = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)\}$ with respect to the hyperplane $(\vec{w}, b)$ is

$$\delta = min_i \;\; y_i(w \cdot x_i + b)$$

Theorem (Novikoff): If for a training set S there exists a weight vector with margin δ, then the perceptron makes at most

$$4\left(\frac{R^2}{\delta^2}\right)$$

mistakes before returning a separating hyperplane.

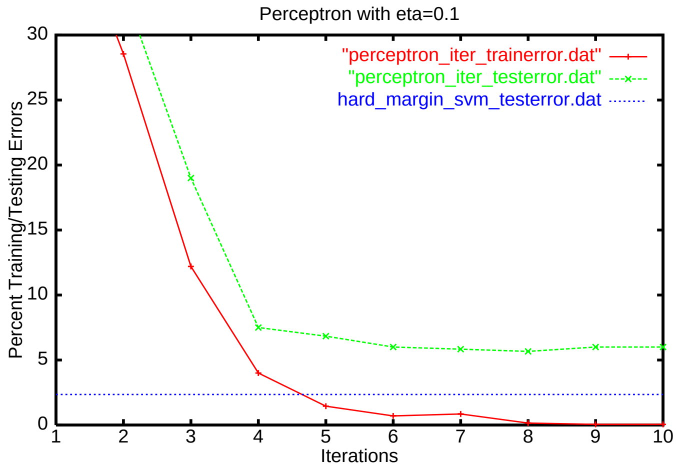
Dual Perceptron

- For each example $(\vec{x}_i, y_i)$, count with α_i the number of times the perceptron algorithm makes a mistake on it. Then

$$\vec{w} = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i$$

```
• $\vec{\alpha} = 0; b_0 \leftarrow 0$ and $R = max_i \|\vec{x}_i\|$
• repeat
    • for i=1 to n
        • if $y_i\left(\sum_{j=1}^{n} \alpha_i y_i (\vec{x}_j \cdot \vec{x}_i) + b_k\right) \le 0$
            • $\vec{\alpha}_i \leftarrow \alpha_i + 1$
            • $b \leftarrow b + y_i R^2$
        • endif
    • endfor
• until no mistakes made in the for loop
• return $(\vec{\alpha}, b)$
```

Experiment: Perceptron for Text Classification



Train on 1000 pos / 1000 neg examples for “ acq” (Reuters-21578).