Categorização de Textos Usando Máquinas de Suporte Vetorial

Vanessa Cristina Sabino

1 de março de 2007

IME - Instituto de Matemática e Estatística

Introdução

- *Problema:* gerenciar eficientemente o conhecimento
- Explosão de informação
    - 2ª Guerra Mundial
    - Vannevar Bush: "A somatória da experiência humana está sendo expandida numa velocidade prodigiosa, e os meios que usamos para achar nosso caminho no labirinto resultante até o item que importa no momento são os mesmos usados nos tempos dos veleiros."
- Capacidade de armazenamento
    - Livro: 100 mil palavras
    - CD-ROM: 65 milhões de palavras
    - DVD: 530 milhões de palavras
    - Internet
- Sobrecarga de informações
    - Vários documentos que cobrem o mesmo tópico

Introdução

- *Jornal:* disseminar notícias, conhecimentos e algum entretenimento
  - Coletar informação de diversas fontes
  - Formato eletrônico e computadores interligados
- Pesquisa
  - Acessar
  - Selecionar
  - Organizar
- Sistemas de busca
  - Regra: ocorrência de determinada palavra
  - Propriedades locais x globais
  - Insuficiente para interpretação semântica

Introdução

- *Solução:* Categorização de Textos
    - Agrupamento de documentos em diferentes categorias ou classes
    - Início: década de 60
    - Década de 80: Criação manual de regras
    - Década de 90: Aprendizagem computacional
        - Processo indutivo
        - Conjunto de documentos previamente classificados

Tarefa de Aprendizado

- Atribuir a uma determinada informação o rótulo da classe a qual ela pertence
- Aprendizagem supervisionada
- Entrada: amostra de treinamendo $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$ i.i.d. de acordo com $Pr(\vec{x}, y_1)$ fixada desconhecida
- Medida de Performance: $R(h) = \int L(h(\vec{x}), y) dPr(\vec{x}, y)$

Classificação Binária

- Duas classes: $y \in \{-1, +1\}$
- Perda 0/1:

$$L_{0/1}(h(\vec{x}), y) = \begin{cases} 0 & h(\vec{x}) = y \\ 1 & \text{caso contrário} \end{cases}$$

- Taxa de erro:
  $Err(h) = Pr(h(\vec{x}) \neq y|h) = \int L_{0/1}(h(\vec{x}), y) dPr(\vec{x}, y)$
- Fatores de custo

Outros Tipos de Classificação

- Múltiplas classes: $y \in \{1, \ldots, n\}$
  - Abordagens não costumam ser computacionalmente eficientes
  - Divisão em n problemas binários
    - $h^{(1)}, \ldots, h^{(n)}$
    - Regra de Bayes: escolhida a classe em que $h^{(i)}(\vec{x})$ é maior
  - Possível reduzir para $l(l-1)/2$ problemas
- Documento em várias categorias
  - Vetor binário n-dimensional $\vec{y} \in \{+1, -1\}^n$
  - Categoria i é destinada a um documento $\vec{x}$ se a regra de classificação correspondente $h^{(1)}(\vec{x})$ resulta $+1$

Representação do Texto

- Representação dos documentos através dos vetores $\vec{x}$
- Problema: Variação de significado de acordo com o contexto
- Abordagens e termos de indexação
  1. Sub-palavra: decomposição das palavras e sua morfologia
  2. **Palavra**: palavras e informação léxica
  3. Multi-palavras: frases e informação sintática
  4. Semântico: significado do texto
  5. Pragmático: contexto e situação (ex.: estrutura de diálogo)

Palavras

- Unidades significativas e pouca ambigüidade
- Simplicidade de implementação

"O vocabulário de uma língua reflete a distribuição à priori das tarefas de classificação de texto: tarefas de classificação de texto para quais o vocabulário contém palavras-chave indicativas são à priori mais prováveis"

Vetores de Documento

- Abordagem *bag-of-words*
- Cada exemplo é um vetor de dimensão fixa
- Cada palavra p é um atributo
- Valor do atributo: número de vezes em que ocorre no documento d: $FT(p, d)$

Seleção de Características

- Etapa de pré-processamento
- Objetivo: eliminar atributos irrelevantes ou inapropriados
- Motivações:
  - Reduzir risco de *overfitting*
  - Aumentar a eficiência computacional em tempo e/ou espaço

Seleção de Subconjuntos de Características

- Eliminação de *stopwords*
- *document frequency thresholding*
  - Conjuntura de Apté e Damerau
- Ganho de informação:

$$\sum_{y \in \{-1,+1\}} \sum_{w \in \{0,1\}} Pr(y,w) \frac{Pr(y,w)}{Pr(y)Pr(w)}$$

- Razão de Chances
- Testes χ^2

Construção de Características

- *stemming*: análise morfológica da palavra e armazena apenas o prefixo
- *tesauros*: abordagem semântica, palavras agrupadas em classes de equivalência
- *indexação semântica latente*: análise de componente principal linear aplicada a textos
- *clusterização de termos*: termos semânticamente similares são agrupados em um cluster gerados através de algoritmos de aprendizado não-supervisionados

Ponderação de Termos

- *componente de documento*: freqüência de termo $FT(p_i, d_j)$
- *componente de coleção*: freqüência de documento $FD(p_i)$
- *componente de normalização*: possibilita que documentos de diferentes tamanhos possam ser comparados na mesma escala

Propriedades das Tarefas de Classificação de Textos

- alta dimensão do espaço de características:
  - Lei de Heaps: $V = ks^{\beta}$, k entre 10 e 100 e β entre 0,4 e 0,6
  - Ex.: Reuters: 9.603 documentos, 27.658 palavras distintas
- vetores esparsos
  - Ex.: Reuters: em média 152 palavras, sendo 74 distintas
- uso de termos heterogêneos
  - "semelhança em família"
- alto nível de redundância
  - maioria dos documentos contém mais de uma palavra que indica a sua classe
- distribuição de freqüência de palavras
  - Lei de Zipf: a n-ésima palavra mais freqüente ocorre $\frac{1}{n}$ vezes a freqüência das palavras mais freqüente
  - Lei de Zipf generalizada: $FT_i = \frac{c}{(k+r)^{\phi}}$

Taxa de Erro e Custo Assimétrico

- Probabilidade da regra de classificação *h* prever a classe errada
$$Err_{teste}(h) = \frac{f_{+-} + f_{-+}}{f_{++} + f_{+-} + f_{-+} + f_{--}}$$
- Matriz de custo ou utilidade $\rightarrow$ função custo linear

Precisão e *Recall*

- Precisão: $Pr(y = 1 | h(\vec{x}) = 1, h)$

$$Prec_{teste}(h) = \frac{f_{++}}{f_{++} + f_{+-}}$$

- *Recall*: $Pr(h(\vec{x}) = 1 | y = 1, h)$,

$$Rec_{teste}(h) = \frac{f_{++}}{f_{++} + f_{-+}}$$

Medida F_β

- média harmônica ponderada entre precisão e *recall*

$$F_\beta(h) = \frac{(1+\beta^2) Prec(h) Rec(h)}{\beta^2 Prec(h) + Rec(h)}$$

$$F_\beta(h) = \frac{(1+\beta^2) f_{++}}{(1+\beta^2) f_{++} + f_{+-} + \beta^2 f_{-+}}$$

Média Micro e Macro

- *média macro*: média aritimética das medidas de performance de cada um dos m experimentos.

$$F_1^{macro} = \frac{1}{m}\sum_{i=1}^{m} F_1(h_i)$$

- *média micro*: obtém-se uma tabela de contingência média

$$F_1^{micro} = \frac{2f_{++}^{med}}{2f_{++}^{med} + f_{+-}^{med} + f_{-+}^{med}}$$

Máquinas de Suporte Vetorial

- *Boa capacidade de generalização*: limites estatísticos para o erro de classificação na população de dados
- *Robustez em grandes dimensões*: não há tendência a *overfitting*
- *Convexidade da função objetivo*: apenas um mínimo global
- *Teoria bem definida*: bem fundamentada em teorias da matemática e estatística.

Aprendizado Estatístico

Objetivo (Teoria de aprendizado estatístico)

Escolha de um classificador f, dentro do conjunto F de todos os classificadores possíveis para aquele conjunto de treinamento S, que seja capaz de classificar dados daquele tipo da forma mais correta possível

- Risco empírico
  - Erro de classificação apenas dentro do conjunto de treinamento
- Risco funcional
  - Probabilidade de que f cometa erro na classificação de um novo exemplo gerado segundo P
  - Desempenho de generalização

Aprendizado Estatístico

- Formalmente:
  - S um conjunto de treinamento
  - Cada exemplo $\vec{x}_i$ pertence ao espaço $\Re^m$
  - Rótulos correspondentes y_i assumem valores -1 ou $+1$
- Objetivo: encontrar uma função $g : \Re^m \rightarrow \{-1, +1\}$ capaz de predizer a classe de novos pontos $(\vec{x}, y)$ de forma precisa
- Solução: função sinal composta com uma função $f(\vec{x})$ que define uma fronteira de separação entre os dados

Limites do Risco Funcional

- número de exemplos de treinamento
- risco empírico obtido neste conjunto
- complexidade do espaço de hipóteses (dimensão VC)

Dimensão Vapnik-Chervonenkis

Definição (Dimensão VC)

Dado um conjunto de funções sinal G, sua dimensão VC é definida como o tamanho do maior conjunto de pontos que pode ser particionado arbitrariamente pelas funções contidas em G

- A dimensão VC de um conjunto de dicotomias G é então definida como a cardinalidade do maior conjunto S que é fragmentado por G, ou seja, o maior N tal que $\Delta_G(S) = 2^N$ em que $N = |S|$.

Limite Superior

Teorema (Limite Superior)

Seja G um conjunto de funções de decisão mapeando $\Re^m$ a $\{-1, +1\}$ com dimensão VC h. Para qualquer distribuição de probabilidade P em $\Re^m \times \{-1, +1\}$, com probabilidade de ao menos $1 - \delta$ sobre n exemplos e para qualquer hipótese g em G o risco funcional é limitado por

$$R(g) \leq R_{emp}(g) + \sqrt{\frac{c}{n}\left(h + ln\left(\frac{1}{\delta}\right)\right)}$$

em que c é uma constante universal. Se $\widehat{g} \in G$ minimiza o risco empírico, então com probabilidade $1 - \delta$

$$R(\widehat{g}) \leq \inf_{g' \in G} R_{emp}(g) + \sqrt{\frac{c}{n}\left(h + ln\left(\frac{1}{\delta}\right)\right)}$$

Limite Superior

- Quanto menor a dimensão VC de uma função, maior sua capacidade de generalização
- Minimização do Risco Estrutural

Conceito de Margem

Definição (Margem)

A margem de um classificador é definida como a menor distância entre os exemplos do conjunto de treinamento e o hiperplano utilizado na separação desses dados em classes.

Conceito de Margem

Teorema

Seja $X_0 \subset \Re^m$ o conjunto de entradas com norma menor que $R > 0$ ($\| \vec{x}_i \| \leq R$, para todo $\vec{x}_i \in X_0$) e F o conjunto de funções lineares definidas em X_0 e satisfazendo $\| f(\vec{x}) \| \geq \rho$, em que ρ é a margem do classificador

$$F = \{\vec{x} \rightarrow \vec{w} \cdot \vec{x} \mid \; \| \vec{w} \| \leq 1, \vec{x} \in X_0\}$$

Considerando G o conjunto de funções sinal obtidas a partir de $G = sgn(F)$ e h a dimensão VC de G, tem-se o resultado

$$h \leq \left\{ \frac{R^2}{\rho^2}, m \right\} + 1$$

Conceito de Margem

- Quanto maior a margem de um classificador, menor sua dimensão VC.

Conceito de Margem

Teorema

Definindo a margem ρ de um classificador f como

$$\rho = min_i y_i f(\vec{x_i}),$$

seja o erro marginal de f ($R_\rho(f)$) a proporção de exemplos de treinamento que tem margem menor que ρ.

$$R_\rho(f) = \frac{1}{n} \sum_{i=1}^{n} |y_i f(\vec{x_i}) < \rho|$$

Conceito de Margem

Teorema (Continuação)

Seja G o conjunto de funções $g(x) = sgn(f(\vec{x})) = sgn(\vec{w} \cdot \vec{x})$ *com* $\| \vec{w} \| \leq \Lambda$ *e* $\| \vec{x} \| \leq R$*, para algum R,* $\Lambda > 0$*. Seja* $\rho > 0$*. Para todas distribuições P gerando os dados, com probabilidade de ao menos* $1 - \delta$ *sobre n exemplos, e para qualquer* $\rho > 0$ *e* $\delta \in (0, 1)$*, a probabilidade de um ponto de teste amostrado independentemente segundo P ser classificado incorretamente é limitado superiormente por*

$$R_\rho(g) + \sqrt{\frac{c}{n}\left(\frac{R^2\Lambda^2}{\rho^2}\ln^2 n + \ln\left(\frac{1}{\rho}\right)\right)}$$

em que c é uma constante universal.

Conceito de Margem

- Minimização do erro
  - sobre os dados de teste: margem ρ alta
  - sobre os dados de treinamento: poucos erros marginais
- Propriedades do hiperplano ótimo:
  - robustez em relação aos padrões
  - robustez em relação aos parâmetros

SVMs Lineares com Margens Rígidas

- Conjunto de treinamento linearmente separável
- Classificador linear: $\vec{w} \cdot \vec{x} + b = 0$

$$\begin{cases} y_i = +1 & \text{se} \quad \vec{w} \cdot \vec{x}_i + b > 0 \\ y_i = -1 & \text{se} \quad \vec{w} \cdot \vec{x}_i + b < 0 \end{cases}$$

Obtenção do Hiperplano Ótimo

- Representação canônica do hiperplano
  - Reescalando $\vec{w}$ e b de forma que os pontos mais próximos do hiperplano separador satisfaçam $|\vec{w} \cdot \vec{x_i} + b| = 1$

$$\begin{cases} \vec{w} \cdot \vec{x_i} + b \geq +1 & \text{se} \quad y_i = +1 \\ \vec{w} \cdot \vec{x_i} + b \leq -1 & \text{se} \quad y_i = -1 \\ i = 1, \ldots, n \end{cases}$$

- Sejam $\vec{x_1}$ e $\vec{x_2}$ pontos sobre as retas $\vec{w} \cdot \vec{x} + b = -1$ e $\vec{w} \cdot \vec{x} + b = +1$, respectivamente, tal que uma reta perpendicular a $\vec{w} \cdot \vec{x_i} + b = 0$ intercepte ambos os pontos.

$$\begin{cases} \vec{w} \cdot \vec{x_1} + b = -1 \\ \vec{w} \cdot \vec{x_2} + b = +1 \end{cases} \Longrightarrow \vec{w} \cdot (\vec{x_2} - \vec{x_1}) = 2$$

Obtenção do Hiperplano Ótimo

- Pela ortogonalidade entre o hiperplano separador e $\vec{w}$ e $\vec{x_2} - \vec{x_1}$
$$|\vec{w} \cdot (\vec{x_2} - \vec{x_1})| = \parallel \vec{w} \parallel \times \parallel \vec{x_2} - \vec{x_1} \parallel$$
- Substituindo
$$\parallel \vec{x_2} - \vec{x_1} \parallel = \frac{2}{\parallel \vec{w} \parallel}$$
- Logo,
  - distância entre os hiperplanos $\vec{w} \cdot \vec{x} + b = 0$ e $\vec{w} \cdot \vec{x} + b = 1$ ou $\vec{w} \cdot \vec{x} + b = -1$ é dada por $\frac{1}{\|\vec{w}\|}$.

Problema de Otimização Quadrática

$$\begin{aligned} \mathit{minimizar}: &\quad \| \vec{w} \|^2 \\ \mathit{sujeito\ a}: &\quad y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 \quad \text{para } i = 1, \ldots, n \end{aligned}$$

Função Lagrangiana

$$L(\vec{w}, b, \alpha) = \frac{1}{2} \parallel \vec{w} \parallel^2 - \sum_{i=1}^{n} \alpha_i (y_i(\vec{w} \cdot \vec{x_i} + b) - 1)$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

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

Problema Dual de Otimização

$$maximizar: \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \vec{x}_i \cdot \vec{x}_j$$

$$sujeito\ a: \quad \begin{cases} \alpha_i \geq 0, \ i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \end{cases}$$

Solução

- α_i^* assume valores positivos para exemplos de treinamento que estão a uma distância do hiperplano ótimo exatamente igual à margem (chamados vetores de suporte) e zero para todos os outros
- $\vec{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \vec{x}_i$
- $b^* = -\frac{1}{2}[\max_{i|y_i=-1}(\vec{w}^* \cdot \vec{x}_i) + \min_{i|y_i=+1}(\vec{w}^* \cdot \vec{x}_i)]$
- Classificação:

$$sgn(\sum_{x_i \in SV} \alpha_i^* y_i \vec{x}_i \cdot \vec{x} + b^*)$$

SVMs Lineares com Margens Suaves

- Problema não linear ou muito ruído nos dados
- Variáveis de relaxamento ξ
  - Medem onde se encontram os exemplos $(\vec{x}_i, y_i)$ em relação aos hiperplanos $\vec{w} \cdot \vec{x} + b = \pm 1$ nos casos em que a classificação está incorreta

$$\text{Para } y_i = +1 \quad \xi_i(\vec{w}, b) = \begin{cases} 0 & \text{se } \vec{w} \cdot \vec{x}_i + b \geq 1 \\ 1 - \vec{w} \cdot \vec{x}_i + b & \text{se } \vec{w} \cdot \vec{x}_i + b < 1 \end{cases}$$

$$\text{Para } y_i = -1 \quad \xi_i(\vec{w}, b) = \begin{cases} 0 & \text{se } \vec{w} \cdot \vec{x}_i + b \leq -1 \\ 1 + \vec{w} \cdot \vec{x}_i + b & \text{se } \vec{w} \cdot \vec{x}_i + b > -1 \end{cases}$$

Problema de Otimização Quadrática

- Variável auxiliar ζ tal que $\zeta_i \geq \xi_i(\vec{w}, b)$

$$
\begin{aligned}
\textit{minimizar}: &\quad \| \vec{w} \|^2 + C \sum_{i=1}^{n} \zeta_i \\
\textit{sujeito a}: &\quad \begin{cases} \zeta_i \geq 0 \\ y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 - \zeta_i \end{cases}
\end{aligned}
$$

Condições de Karush-Kuhn-Tucker

- $\alpha_i = 0 \Rightarrow y_i f(\vec{x_i}) \geq 1 \text{ e } \zeta_i = 0$
- $0 < \alpha_i < C \Rightarrow y_i f(\vec{x_i}) = 1 \text{ e } \zeta_i = 0$
- $\alpha_i = C \Rightarrow y_i f(\vec{x_i}) < 1 \text{ e } \zeta_i \geq 0$

SVMs Não-Lineares

- Funções reais $\Phi_1, \ldots, \Phi_M$ que mapeiam o conjunto de treinamento S para o *espaço de características* de forma a torná-lo linearmente separável

$$\vec{\Phi}(S) = \{(\vec{\Phi}(\vec{x_1}), y_1), \ldots, (\vec{\Phi}(\vec{x_n}), y_n)\}$$

- Basta saber calcular o produto interno $\vec{\Phi}(\vec{x_i}) \cdot \vec{\Phi}(\vec{x_j})$

$$K(x, z) = \Phi(x) \cdot \Phi(z)$$

- Teorema de Mercer: os kerneis devem ser matrizes positivas semi-definidas para qualquer subconjunto finito de S

Principais Kerneis

Tipo de Kernel	**Função** $K(\vec{x_i}, \vec{x_j})$	**Comentários**
Polinomial	$(\vec{x_i} \cdot \vec{x_j} + 1)^p$	A potência p deve ser especificada pelo usuário
Gaussiano	$e^{(-\frac{1}{2\sigma^2}\|\vec{x_i} - \vec{x_j}\|^2)}$	A ampliture σ^2 é especificada pelo usuário
Sigmoidal	$tanh(\beta_0(\vec{x_i} \cdot \vec{x_j}) + \beta_1)$	Utilizado somente para alguns valores de β_0 e β_1

SVMs Incrementais

- Vantagens:
  - tornar os aprendizados sucessivos mais rápidos
  - reduzir o custo de armazenamento descartando exemplos

Algoritmo α-ISVM

- 3 conceitos:
    - conjunto de backup: exemplos que nunca são selecionados como vetores de suporte (*intra-sample*)
    - conjunto de cache: exemplos que freqüentemente aparecem como vetores de suporte (*vice-boundary sample*)
    - conjunto de trabalho: exemplos do último conjunto de vetores de suporte (*boundary sample*)

Algoritmo α-ISVM

- Método Iterativo:
  1. classificador antigo é utilizado no novo conjunto de exemplos incremental
     - aqueles que forem classificados incorretamente são combinados ao conjunto de vetores de suporte atual para construir um novo conjunto de treinamento,
     - os outros exemplos formam um novo conjunto de testes.
  2. novo classificador é treinado no novo conjunto de treinamento
  3. novo conjunto de testes é utilizado para repetir a operação anterior

Algoritmo α-ISVM

- Medidas para reduzir o custo de armazenagem e acelerar a convergência:
  - exemplos do conjunto de backup são descartados gradualmente usando o esquema LRU
  - exemplos do conjunto de trabalho são diretamente introduzidos no conjunto de treinamento
  - exemplos do conjunto de cache são introduzidos no conjunto de treinamento de acordo com uma certa prioridade

Limitando o Erro Esperado Baseado na Margem

Teorema (Limite no Erro Esperado de SVMs de Margens Suaves)

O erro esperado $\varepsilon(Err^n(h_{SVM}))$ *de um SVM de margem suave baseado em n exemplos de treinamento com* $c \leq K(\vec{x}_i, \vec{x}_j) \leq c + R^2$ *para alguma constante c é limitado por*

$$\varepsilon(Err^n(h_{SVM})) \leq \frac{\rho\varepsilon(\frac{R^2}{\delta^2}) + \rho C'\varepsilon\left(\sum_{i=1}^{n+1} \xi_i\right)}{n+1}$$

com $C' = CR^2$ *se* $C \geq \frac{1}{\rho R^2}$, *e* $C' = CR^2 + 1$ *caso contrário.*

Limitando o Erro Esperado Baseado na Margem

- margem δ
- perda de treinamento ξ
- quantidade R associada ao tamanho dos vetores de documento, que atua como uma constante para escalar a margem δ.

Conceitos TCat Homogêneos

Definição (Conceitos TCat Homogêneos)

O conceito TCat

$$TCat([p_1 : n_1 : f_1], \ldots, [p_s : n_s : f_s])$$

descreve uma tarefa de classificação binária com s conjuntos disjuntos de características. O i-ésimo conjunto inclui f_i características. Cada exemplo positivo contém p_i ocorrências de características do conjunto respectivo, e cada exemplo negativo contém n_i ocorrências. Uma mesma característica pode ocorrer múltiplas vezes em um documento.

Conceitos TCat Homogêneos

- Exemplo:

  TCat([20 : 20 : 100], [4 : 1 : 200], [1 : 4 : 200], [5 : 5 : 600],
  [9 : 1 : 3000], [1 : 9 : 3000], [10 : 10 : 4000])
- Propriedades
  - alta dimensão do espaço de entrada
  - vetor de documento esparso
  - alto nível de redundância
  - uso heterogêneo de termos
  - Lei de Zipf

Conceitos TCat Homogêneos

$$h(\vec{x}) = \vec{w} \cdot \vec{x} + b = \sum_{i=1}^{11100} w_i x_i + b$$

com $b = 0$ e

$$w_i = \begin{cases} +0.23 & \text{para as 200 palavras de média freqüência indicando POS} \\ -0.23 & \text{para as 200 palavras de média freqüência indicando NEG} \\ +0.04 & \text{para as 3000 palavras de baixa freqüência indicando POS} \\ -0.04 & \text{para as 3000 palavras de baixa freqüência indicando NEG} \\ 0 & \text{para todas as outras palavras} \end{cases}$$

$\Rightarrow$ margem δ: $\sqrt{1/30,15}$

Capacidade de Aprendizagem de Conceitos TCat

Lema (Limite inferior da margem de conceitos TCat livres de ruído)

Para um conceito $TCat([p_1 : n_1 : f_1], \ldots, [p_s : n_s : f_s])$*, existe sempre um hiperplano passando através da origem que tem margem* δ *limitada por*

$$\delta^2 \geq \frac{ac - b^2}{a + 2b + c} \qquad com \qquad a = \sum_{i=1}^{s} \frac{p_i^2}{f_i} \quad b = \sum_{i=1}^{s} \frac{p_i n_i}{f_i} \quad c = \sum_{i=1}^{s} \frac{n_i^2}{f_i}$$

Capacidade de Aprendizagem de Conceitos TCat

Lema (Distância Euclidiana dos Vetores de Documento)

Se as freqüências de termos rankeadas FT_r em um documento com l termos têm a forma da Lei de Zipf generalizada

$$TF_r = \frac{c}{(r+k)^{\phi}}$$

baseado em seu rank de freqüência r, então o quadrado da distância euclidiana do vetor de documento $\vec{x}$ de freqüências de termos é limitado por

$$\| \vec{x} \| \leq \sqrt{\sum_{r=1}^{d} \left(\frac{c}{(r+k)^{\phi}} \right)^2} \qquad com\ d\ tal\ que \qquad \sum_{r=1}^{d} \frac{c}{(r+k)^{\phi}} = l$$

Capacidade de Aprendizagem de Conceitos TCat

- Devido a Lei de Zipf, a distância euclidiana é menor do que l, pois a maioria dos termos não se repete muito freqüentemente e o número de termos distintos d é alto. Isso leva a um valor baixo de R^2 no limite na performance de generalização esperada.

Capacidade de Aprendizagem de Conceitos TCat

Teorema (Capacidade de Aprendizagem de Conceitos TCat)

Para conceitos

$$TCat([p_1 : n_1 : f_1], \ldots, [p_s : n_s : f_s])$$

e documentos com l termos distribuídos de acordo com a Lei de Zipf generalizada

$$TF_r = \frac{c}{(r+k)^\phi},$$

o erro de generalização esperado de uma SVM após treinamento em n exemplos é limitado por

Capacidade de Aprendizagem de Conceitos TCat

Teorema (Continuação)

$$\varepsilon(Err^n(h_{SVM})) \leq \rho \frac{R^2}{n+1} \frac{ac - b^2}{a + 2b + c} \; com \quad \begin{aligned} a &= \sum_{i=1}^{s} \frac{p_i^2}{f_i} \\ b &= \sum_{i=1}^{s} \frac{p_i n_i}{f_i} \\ c &= \sum_{i=1}^{s} \frac{n_i^2}{f_i} \\ R^2 &= \sum_{i=1}^{s} \left(\frac{c}{(r+k)^{\phi}} \right)^2 \end{aligned}$$

a não ser que $\forall_{i=1}^{s} : p_i = n_i$.
d é escolhido tal que $\sum_{r=1}^{d} \frac{c}{(r+k)^{\phi}} = l$.

Resultados - Taxa de Erro

	modelo	experimento
WebKB “course”	11,2%	4,4%
Reuters “earn”	1,5%	1,3%
Ohsumed “pathology”	94,5%	23,1%

Classificador Naive Bayes

$$h_{BAYES}(d) = argmax_{y \in \{-1,+1\}} \frac{Pr(y) \cdot \prod_{i=1}^{|d|} Pr(w_i|y)}{\sum_{y' \in \{-1,+1\}} Pr(y') \cdot \prod_{i=1}^{|d|} Pr(w_i|y')}$$

Algortimo Rocchio

$$\vec{w} = \frac{1}{|i : y_i = +1|} \sum_{i:y_i=+1} \vec{x}_i - \beta \frac{1}{j : y_j = -1} \sum_{|j:y_j=-1|} \vec{x}_j$$

- Para classificar um novo documento, é calculado o cosseno entre $\vec{x}$ e $\vec{w}$ como medida de similaridade

k-nearest neighbours

$$h_{knn}\vec{x} = sign\left(\frac{\sum_{i \in knn(\vec{x})} y_i cos(\vec{x}, \vec{x_i})}{\sum_{i \in knn(\vec{x})} cos(\vec{x}, \vec{x_i})}\right)$$

Outros métodos

- Classificador de Árvore de Decisão (C4.5)
- Rede Bayesiana
- Regressão Logística
- Redes Neurais
- Regressão Polinomial (regressão linear)
- Algoritmos de Boosting (AdaBoost)
- Aprendizagem de Regras (busca genética)
- Aprendizagem de Regras Relacional
- Aprendizagem Ativa

FIM