Máquinas de suporte vetorial e sua aplicação na detecção de spam

Antonio Carlos dos Santos
Orientador: Paulo J. S. Silva (IME-USP)

Universidade de São Paulo
Instituto de Matemática e Estatística
Departamento de Ciência da Computação

MAC499 Trabalho de Formatura Supervisionado 2005

Conteúdo

Informações sobre spams

Spams (***S**tupid **P**ointless **A**nnoying **M**essages*): e-mails indesejados que as pessoas recebem.

- 31 milhões de e-mails enviados diariamente em 2004 [Sfr05]
- 12,4 milhões eram spams [Sfr05]
- Perdas mundiais chegarão a US$50 bilhões em 2005, principalmente devido à queda de produtividade dos funcionários [Fer05]

Informações sobre spams

Spams (***S**tupid **P**ointless **A**nnoying **M**essages*): e-mails indesejados que as pessoas recebem.

- 31 milhões de e-mails enviados diariamente em 2004 [Sfr05]
- 12,4 milhões eram spams [Sfr05]
- Perdas mundiais chegarão a US$50 bilhões em 2005, principalmente devido à queda de produtividade dos funcionários [Fer05]

Informações sobre spams

Spams (***S**tupid **P**ointless **A**nnoying **M**essages*): e-mails indesejados que as pessoas recebem.

- 31 milhões de e-mails enviados diariamente em 2004 [Sfr05]
- 12,4 milhões eram spams [Sfr05]
- Perdas mundiais chegarão a US$50 bilhões em 2005, principalmente devido à queda de produtividade dos funcionários [Fer05]

Informações sobre spams

Spams (***S**tupid **P**ointless **A**nnoying **M**essages*): e-mails indesejados que as pessoas recebem.

- 31 milhões de e-mails enviados diariamente em 2004 [Sfr05]
- 12,4 milhões eram spams [Sfr05]
- Perdas mundiais chegarão a US$50 bilhões em 2005, principalmente devido à queda de produtividade dos funcionários [Fer05]

Problemas com spams

- Tempo gasto analisando mensagens para decidir se é spam ou não
- Risco de apagar mensagens que não são spams
- Prejuízo por vírus e outros programas maliciosos

Pergunta:

- Por que não fazemos um programa para classificar automaticamente os spams?
- Dificuldades:
  - Caracterizar um spam
  - Acompanhar o surgimento de novos formatos de spam

Pergunta:

- Por que não fazemos um programa para classificar automaticamente os spams?
- Dificuldades:
  - Caracterizar um spam
  - Acompanhar o surgimento de novos formatos de spam

Idéia

- Criarmos um programa que classifique as mensagens como spam ou não através da análise prévia de algumas mensagens (exemplos) que já tenham sido analisadas, semelhante a como aprendemos a fazer algumas tarefas

Conceitos

- Mundo real: conjunto no qual estamos interessados
- Dado ou exemplo: elemento do mundo real
- Máquina de aprendizado: dispositivo (um programa) capaz de analisar dados do mundo real e classificá-los (dar rótulos a eles).
- Conceito: a classificação correta dos dados que a máquina tenta aprender
- Hipótese: cada uma das possíveis classificações que a máquina faz sobre os dados

Objetivo

- Fazer com que a máquina de aprendizado gere uma hipótese que melhor aproxima o conceito desejado
- Para isso, devemos treinar a máquina de aprendizado com exemplos já classificados corretamente

Fases

1. Treinamento: definir o padrão de classificação
2. Teste: classificar novos dados

Máquinas de suporte vetorial
Apresentação

- Desenvolvidas principalmente por V. Vapnik, usam idéias de aprendizado estatístico
- Têm algumas vantagens sobre outros métodos de classificação
  - Sem mínimos locais
  - Fase de testes é rápida

Máquinas de suporte vetorial

Sejam l dados de treinamento (amostras), cada um formado por um vetor $\boldsymbol{x_i} \in \mathbb{R}^d$ e um rótulo $y_i \in \{-1, 1\}$

- Os dados possuem uma distribuição de probabilidade $P(\boldsymbol{X}, Y)$ desconhecida e são independentes e identicamente distribuídos
- Mas iremos supor que os rótulos são fixos

Máquinas de suporte vetorial

Sejam l dados de treinamento (amostras), cada um formado por um vetor $\boldsymbol{x_i} \in \mathbb{R}^d$ e um rótulo $y_i \in \{-1, 1\}$

- Os dados possuem uma distribuição de probabilidade $P(\boldsymbol{X}, Y)$ desconhecida e são independentes e identicamente distribuídos
- Mas iremos supor que os rótulos são fixos

Máquinas de suporte vetorial

Sejam l dados de treinamento (amostras), cada um formado por um vetor $\boldsymbol{x_i} \in \mathbb{R}^d$ e um rótulo $y_i \in \{-1, 1\}$

- Os dados possuem uma distribuição de probabilidade $P(\boldsymbol{X}, Y)$ desconhecida e são independentes e identicamente distribuídos
- Mas iremos supor que os rótulos são fixos

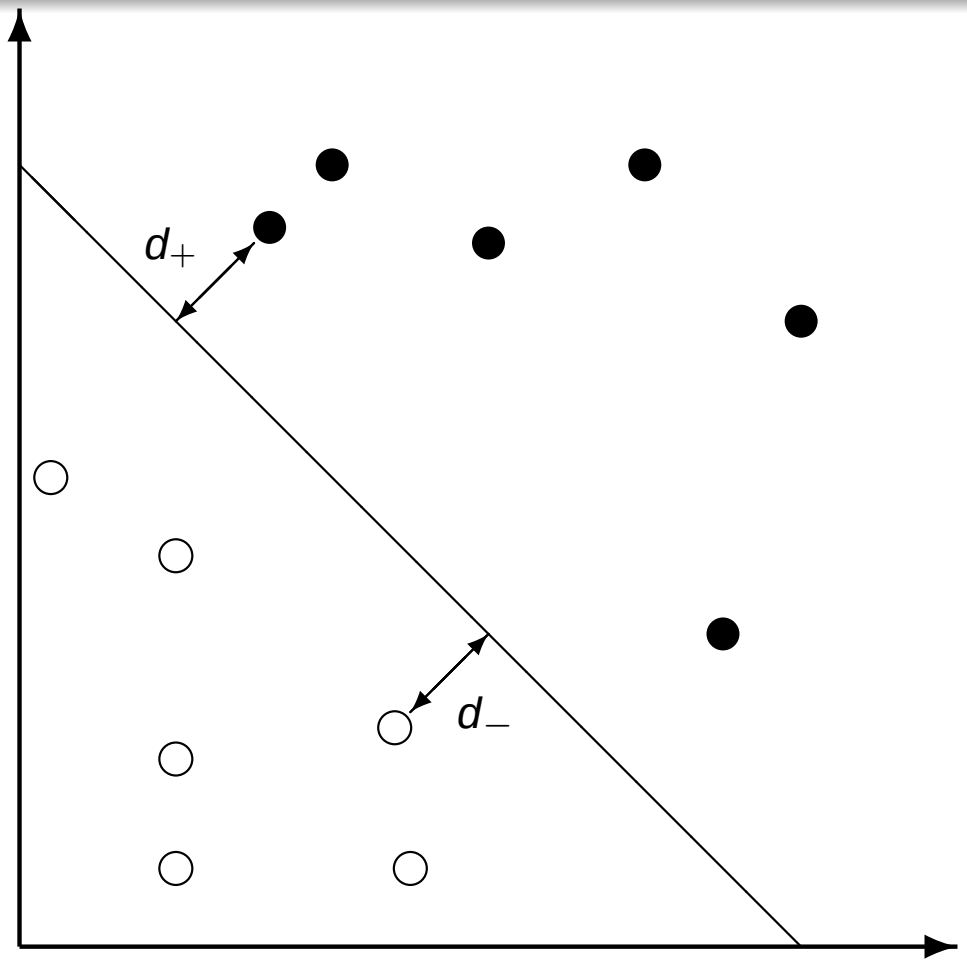
Modelos

- Dados linearmente separáveis
- Dados não-linearmente separáveis

Hiperplano separador

- Um hiperplano em $\mathbb{R}^d$ que separa as amostras positivas ($y_i = +1$) das negativas ($y_i = -1$)
- $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$ é uma equação para este hiperplano
- $\frac{|b|}{\|\boldsymbol{w}\|}$ é a distância perpendicular do hiperplano da origem
- d_+ será a menor distância dos pontos classificados como positivos ao hiperplano e d_- será a menor distância dos pontos classificados como negativos ao hiperplano
- Definiremos a margem por $(d_+ + d_-)$

Exemplo



Hiperplano separador

- Mudando a escala entre $|b|$ e $\|\boldsymbol{w}\|$, teremos:

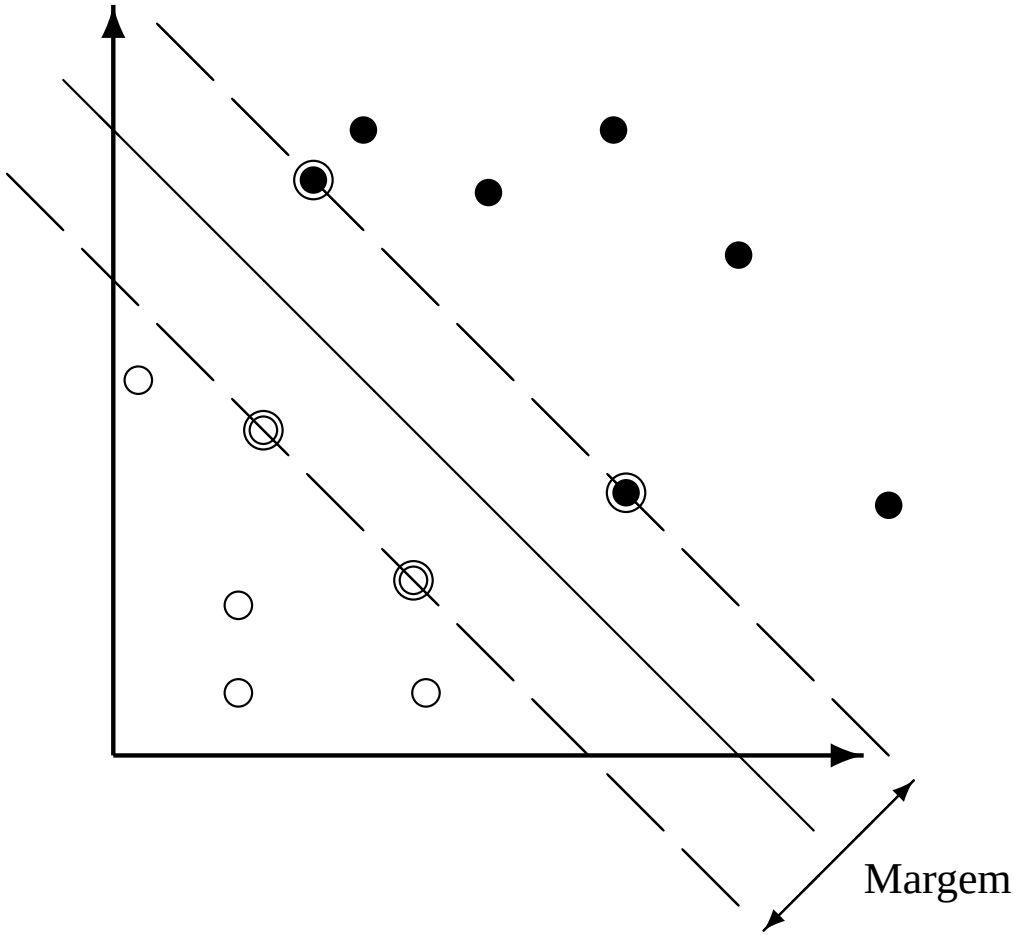
$$\left.\begin{array}{lll} \langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b \geq +1 & \text{se} & y_i = +1 \\ \langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b \leq -1 & \text{se} & y_i = -1 \end{array}\right\} \qquad i = 1, \ldots, l.$$

- Podemos combinar as duas restrições em uma só:

$$y_i(\langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b) - 1 \geq 0 \qquad i = 1, \ldots, l.$$

- Vetores suporte: dados de treinamento para os quais é válida a igualdade nas restrições acima
- A margem será $\frac{2}{\|\boldsymbol{w}\|}$

Exemplo



Objetivo

- Máquinas de suporte vetorial procuram maximizar a margem do hiperplano separador
- Portanto, temos o seguinte problema:

$$\begin{aligned} \max \quad & 2/\|\boldsymbol{w}\| \\ \text{s.a} \quad & y_i(\langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b) - 1 \geq 0 \qquad i = 1, \ldots, l. \end{aligned}$$

Objetivo

- Máquinas de suporte vetorial procuram maximizar a margem do hiperplano separador
- Portanto, temos o seguinte problema:

$$\begin{aligned} \max \quad & 2/\|\boldsymbol{w}\| \\ \text{s.a} \quad & y_i(\langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b) - 1 \geq 0 \qquad i = 1, \ldots, l. \end{aligned}$$

Problema

- Sendo $\|\boldsymbol{w}\| \geq 0$, o problema é equivalente a:

$$\begin{array}{ll} \min & \|\boldsymbol{w}\|^2/2 \\ \text{s.a} & y_i(\langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b) - 1 \geq 0 \qquad i = 1, \ldots, l. \end{array}$$

- A função Lagrangeana associada é:

$$\begin{aligned} L_P(\boldsymbol{w}, b, \boldsymbol{\alpha}) &= \frac{\|\boldsymbol{w}\|^2}{2} - \sum_{i=1}^{l} \alpha_i y_i(\langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b) + \sum_{i=1}^{l} \alpha_i \\ \alpha_i &\geq 0, \quad i = 1, \ldots, l. \end{aligned}$$

em que $\boldsymbol{\alpha}$ é o vetor dos multipilicadores de Lagrange

Problema dual

- A partir das condições KKT, temos o seguinte problema dual equivalente:

$$\begin{aligned} \max \quad & L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{l} \alpha_i \; - \; \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle \\ \text{s.a} \quad & \alpha_i \geq 0 \qquad i = 1, \ldots, l. \end{aligned}$$

no qual os dados de treinamento só aparecem como produtos internos.

Classificação de novos dados

- Após termos treinado a máquina, como sabemos a classe de um novo dado?
- Basta ver em qual lado do hiperplano o dado está:

$$\operatorname{sgn}(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) = \operatorname{sgn}\left(\sum_{i=1}^{l} \alpha_i y_i \langle \boldsymbol{x_i}, \boldsymbol{x} \rangle \right)$$

Dados não-linearmente separáveis

- Para dados não linearmente separáveis, nós introduzimos uma variável de folga $\xi_i \geq 0$ na função objetivo para cada dado de treinamento, de forma a penalizar a violação das restrições do problema original
- Novo problema:

$$\begin{aligned} \min \quad & \frac{\|\boldsymbol{w}\|^2}{2} + C\sum_{i=1}^{l}\xi_i \\ \text{s.a} \quad & y_i(\langle \boldsymbol{x_i}, \boldsymbol{w}\rangle + b) - 1 + \xi_i \geq 0 \quad i = 1,\ldots,l \\ & \xi_i \geq 0 \qquad\qquad\qquad\qquad\quad i = 1,\ldots,l, \end{aligned}$$

em que C e k são parâmetros escolhidos pelo usuário e determinam a importância dos erros de classificação feitos pelo algoritmo.

Aplicação na detecção de spam

- O estudo foi baseado principalmente no artigo *Support Vector Machines for Spam Categorization* [HWV99]
- Característica: uma palavra de um e-mail
- A cada mensagem, teremos um vetor de características $\boldsymbol{x}$ associado, formado por palavras de um dicionário gerado pela análise de todas as mensagens.
- Apenas palavras que aparecem em pelo menos 3 e-mails diferentes

Abordagens

- Frequência da palavra: número que cada palavra do dicionário ocorre no texto
- Representação binária: se cada palavra ocorre ou não no texto

Foram usadas todas as palavras ao invés de selecionarmos apenas algumas

BCC

Disciplinas importantes:

- MAC0122 Princípios de Desenvolvimento de Algoritmos
- MAC0338 Análise de Algoritmos
- MAC0315 Programação Linear
- MAC0300 Métodos Numéricos de Álgebra Linear
- MAC5732 Aprendizado Computacional
- MAT0121 Cálculo Diferencial e Integral II
- MAT0139 Álgebra Linear para Computação
- MAE0212 Introdução à Probabilidade e à Estatística II

Para quem quiser saber mais: I


C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery,* 2:121-167, 1998.


N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based methods.* Cambridge University Press, 2002.


Ferris Research. http://www.ferris.com, Outubro de 2005.


H. Druker, D. Wu, V. N. Vapnik. Support Vector Machines for Spam Categorization. *IEEE Transactions on Neural Networks,* 10(5):1048-1054, 1999.

Para quem quiser saber mais: II


Spam Filter Review. http://spam-filter-review.toptenreviews.com, Outubro de 2005.


V. N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.