Máquinas de suporte vetorial e sua aplicação na detecção de spam

Antonio Carlos dos Santos **NºUSP: 3534575** **Orientador: Paulo J. S. Silva - IME - USP**

acsantos@linux.ime.usp.br

Introdução

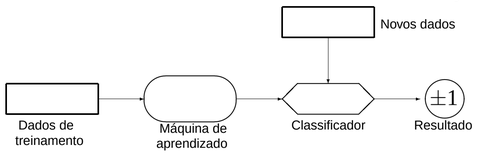
Com o uso cada vez maior da Internet, a troca de mensagens eletrônicas, os *e-mails*, tornou-se uma ação muito frequente entre seus usuários, tanto para assuntos profissionais quanto para pessoais. Entretanto, o número de mensagens indesejadas recebidas, os **spams** (*Stupid Pointless Annoying Messages*), também é muito grande. Esta situação tem gerado vários problemas, tanto para empresas provedoras de acesso à Internet, que têm uma carga maior de uso de seus servidores, quanto para os usuários, que gastam muito tempo lendo spams para depois descartá-los, e podem ainda serem prejudicados por programas maliciosos, como vírus e *spywares*.
Assim, para evitar tais problemas, atualmente tem-se pesquisado bastante novas formas de detectar e bloquear mensagens consideradas spams automaticamente. Entre elas, está o uso de máquinas de suporte vetorial.
Máquinas de suporte vetorial, em inglês *Support Vector Machines* (SVMs), representam um conceito novo na área de sistemas de aprendizado computacional. São baseadas na teoria de aprendizado estatístico, desenvolvida principalmente por Vladimir Vapnik[Vap98], tendo com idéia principal o mapeamento do dados de entrada para um outro espaço onde haja um hiperplano que os separe linearmente. SVMs têm apresentado um bom desempenho em algumas aplicações como, por exemplo, reconhecimento de imagens e, além disso, têm como uma vantagem sobre redes neurais o fato de não apresentarem mínimos locais.

Aprendizado computacional

Sejam l dados de treinamento (amostras), cada um formado por um vetor $\boldsymbol{x_i} \in \mathbb{R}^d$ e um rótulo $y_i \in \{-1, 1\}$. Suponhamos que estes dados possuam uma distribuição de probabilidade $P(\boldsymbol{X}, Y)$ desconhecida e sejam independentes e identicamente distribuídos.
Uma *máquina de aprendizado computacional* (no nosso caso, um programa) será encarregada de aprender o mapeamento:

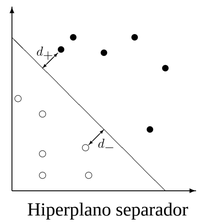
$$\boldsymbol{x_i} \longmapsto y_i, \quad i = 1, \cdots, l.$$

E, para cada novo dado de teste, o programa o classificará de acordo com este mapeamento, tentando reduzir os erros de classificação.



Máquinas de suporte vetorial

Suponhamos inicialmente que nossos dados de treinamento sejam linearmente separáveis, ou seja, que há um hiperplano $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$ tal que todos os dados de treinamento classificados como positivos ($y_i = +1$) fiquem de um lado do hiperplano e os dados classificados como negativos ($y_i = -1$) fiquem do outro lado do mesmo hiperplano. Sejam d_+ a distância perpendicular do dado de treinamento positivo mais próximo ao hiperplano e d_- a distância do dado negativo mais próximo ao hiperplano:



Hiperplano separador

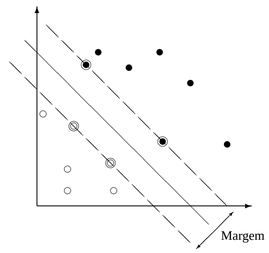
Definiremos a *margem* de um hiperplano por $d_+ + d_-$. Máquinas de suporte vetorial procuram maximizar a margem do hiperplano separador. Mudando a escala entre $\|w\|$ e $|b|$, podemos supor que:

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

Da equação acima, observamos que o hiperplano separador tem uma margem igual a $2/\|\boldsymbol{w}\|$. Assim, temos o seguinte problema de otimização:

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

Os pontos para os quais vale a igualdade em 1 são chamados *vetores suporte*, em inglês *support vectors (sv)*. Tais pontos definem o hiperplano separador e sua remoção ou deslocamento pode afetar a solução do problema 2, pois pode mudar o valor da margem. A figura abaixo ilustra o problema para duas dimensões, com os vetores suporte associados:



O problema lagrangeano dual associado ao problema original é:

$$\max \quad L_D(\boldsymbol{w}, b, \alpha) = \sum_{i=1}^{l} \alpha_i \ - \ \frac{1}{2} \sum_{i=1}^{l} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle$$

$$\text{s. a} \quad \sum_{i=1}^{l} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, l$$

em que $\boldsymbol{\alpha}$ é o vetor dos multiplicadores de Lagrange associados às restrições do problema original. O valor de $\boldsymbol{w}$ será dado por:

$$\boldsymbol{w} = \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x_i}.$$

Após a fase de treinamento, a classe de um novo dado $\boldsymbol{x}$ será determinada por:

$$\text{sgn}(\langle \boldsymbol{x}, \boldsymbol{w} \rangle + b).$$

Assim, a classificação de um novo dado é rápida, pois é preciso computar apenas o produto interno entre $\boldsymbol{w}$ e $\boldsymbol{x}$.
Para dados não-linearmente separáveis, nós adicionamos variáveis de folga ξ_i, de forma a penalizar o valor da função objetivo original quando houver uma violação das restrições 1. O novo problema é:

$$\begin{aligned} &\min \qquad \|w\|^2/2 + C \sum_{i=1}^{l} \xi_i \\ &\text{s.a} \ \ y_i(\langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b - 1 + \xi_i) \ \geq \ 0 \qquad i = 1, \ldots, l \\ &\qquad\qquad\qquad\qquad \xi_i \ \geq \ 0 \qquad i = 1, \ldots, l, \end{aligned} \tag{3}$$

em que C é um parâmetro escolhido pelo usuário e define a gravidade da violação das restrições, ou seja, dos erros de treinamento.

Aplicação na detecção de spam

O estudo da aplicação de SVMs na detecção de spam foi baseado principalmente no artigo *Support Vector Machines for Spam Categorization* [HWV99], em que é feita a comparação de desempenho de SVMs com outros técnicas de classificação.
Neste problema, uma *característica* será definida como 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.
Foram escolhidas duas abordagens para a criação de $\boldsymbol{x}$:

- **Frequência da palavra:** a i-ésima coordenada do vetor de caractéristicas indica o número de vezes que a i-ésima palavra do dicionário aparece na mensagem.
- **Representação binária:** a i-ésima coordenada do vetor de características indica se a i-ésima palavra do dicionário está presente ou não na mensagem.

Em ambas as abordagens, uma palavra pertencerá ao dicionário somente se aparecer em, pelo menos, 3 e-mails diferentes. Tal técnica foi adotada para eliminar mensagens com erros de digitação e para reduzir o tamanho do dicionário, sem afetar significativamente a classificação.
Preferimos usar todas as palavras (características), ao invés de selecionar algumas apenas, porque tal abordagem é mais rápida (o consumo tempo da seleção é quadrático) e, em SVMs, o uso de todas as caracterticas tem apresentado um melhor desempenho.

Referências

[HWV99] H. Druker, D. Wu, V. N. Vapnik. Support Vector Machines for Spam Categorization. *IEEE Transactions on Neural Networks,* 10(5):1048-1054, 1999.

[Vap98] V. N. Vapnik. *Statistical Learning Theory.* John Wiley & Sons, 1998.