Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 8, n. 1, 2021.

Trabalho apresentado no XL CNMAC, Evento Virtual - Co-organizado pela Universidade do Mato Grosso do Sul (UFMS).

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Otimização e análise teórica das máquinas de vetores suporte aplicadas à classificação de dados

Paula Cristina Rohr Ertel[1]
Bolsista PIBIC/CNPq UFSC (2019 -2020)
Luiz Rafael dos Santos[2]
Orientador, UFSC, Campus Blumenau, SC

Em problemas que exigem a análise de uma grande quantidade de dados para classificá-los um processo manual torna-se inviável, motivando o desenvolvimento de técnicas computacionais capazes de reconhecer padrões para desempenhar tal tarefa. Assim, o objetivo central deste trabalho de Iniciação Científica foi desenvolver um estudo teórico, do ponto de vista da otimização, de uma técnica de aprendizagem de máquina supervisionada aplicada à classificação binária de dados denominada Máquinas de Vetores Suporte (SVMs, do inglês *Support Vector Machines*).

A Aprendizagem de Máquina (ML, do inglês *Machine Learning*) é um campo da inteligência computacional que estuda o uso de técnicas computacionais para automaticamente detectar padrões em dados e utilizá-los para fazer predições e tomar decisões. A técnica SVM é uma técnica de ML empregada em problemas de regressão e de classificação, sendo caracterizada como uma técnica de aprendizado supervisionado, pois se utiliza de um conjunto de dados cujas saídas são previamente conhecidas para detectar padrões e produzir um modelo capaz de deduzir as saídas corretas para novos dados. Tal técnica é fundamentada na Teoria de Aprendizagem Estatística e foi desenvolvida por Vladimir Vapnik, Bernhard Boser, Isabelle Guyon e Corrina Cortes [1, 2].

Dependendo do conjunto de dados e da complexidade do problema, a técnica SVM apresenta três diferentes formulações: SVM com margem rígida, SVM com margem flexível e SVM não-linear. No caso em que os dados são linearmente separáveis, isto é, existe um hiperplano que separa corretamente tais dados, realizamos a modelagem matemática da técnica SVM com margem rígida, que resulta no seguinte problema de programação quadrática convexa e com restrições lineares

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1, \quad i = 1, \ldots, m, \end{aligned} \tag{1}$$

em que $w \in \mathbb{R}^n$ e $b \in \mathbb{R}$.

Para problemas cujos dados não são linearmente separáveis podemos formular a técnica SVM de margem flexível, em que, promovendo um relaxamento das restrições através de variáveis de folga ξ_i associadas a cada atributo x^i, ainda é possível obter um hiperplano que nos forneça uma boa classificação. Neste contexto, o problema de encontrar o hiperplano ótimo (1) pode ser reformulado da seguinte maneira

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\ \text{s.a.} \quad & y_i(w^T x^i + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m, \\ & \xi_i \geq 0, \quad i = 1, \ldots, m, \end{aligned} \tag{2}$$

[1] paulaertel@ime.usp.br
[2] l.r.santos@ufsc.br

em que o parâmetro de penalização $C > 0$ tem como objetivo controlar a importância das variáveis de folga ao minimizar a função objetivo.

A técnica SVM não linear é aplicada em problemas cujos dados não são linearmente separáveis e a SVM com margem flexível não fornece uma boa classificação. Na modelagem dessa técnica o hiperplano ótimo é obtido através de um mapeamento dos dados para um espaço de dimensão elevada [6]. É importante salientar que neste trabalho abordamos somente os casos de SVM com margem rígida e flexível.

Em resumo, a formulação matemática da técnica SVM para classificação binária se concentra em obter um hiperplano que melhor separa os dados em duas classes, de modo a possibilitar a máxima margem de separação. A partir disso, derivamos o problema de otimização com restrições lineares cuja solução (w^*, b^*) fornece o hiperplano separador definido pela equação ${w^*}^T x + b^* = 0$ que atuará como classificador para novos dados.

Tendo em vista que os problemas (1) e (2) consistem em problemas de programação quadrática convexa com restrições lineares, abordamos aspectos da teoria de otimização, com e sem restrições, assim como apresentamos definições e resultados de otimização convexa, as quais fornecem propriedades importantes relacionadas aos problemas de otimização, como a garantia de existência de soluções. Ademais, demonstramos que o problema de otimização decorrente da modelagem matemática da técnica SVM com margem rígida admite uma única solução.

Por fim, utilizando a linguagem de programação Julia, realizamos uma implementação computacional da técnica SVM para classificar o conjunto de dados Flor Íris [4] em relação às suas espécies e, posteriormente, para classificar um conjunto de dados sobre células de câncer de mama em tumor maligno ou benigno [7]. Através desses experimentos numéricos foi possível analisar a eficiência da técnica SVM, que apresentou uma classificação 100% correta para os dados do conjunto Íris e atingiu uma acurácia de 96,18% na classificação dos dados de câncer de mama. Em particular, no caso em que aplicamos SVM com margem flexível, tal eficiência está relacionada com a escolha de um parâmetro de penalização adequado.

Referências

[1] Boser, B. E., Guyon, I. M., Vapnik, V. A Training Algorithm for Optimal Margin Classifiers, *Proceedings of the fifth annual workshop on Computational learning theory*, ACM, p. 144–152, 1992.

[2] Cortes, C., Vapnik, V. Support - Vector Networks, *Machine Learning, Springer*, 20.3, p. 273–297, 1995.

[3] Deisenroth, P., Faisal, A. A. and Ong, C. S. *Mathematics for Machine Learning*. Cambridge University Press, Boston, 2019.

[4] Fisher, R. A. The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7.2, p. 179–188, 1936. DOI:10.1111/j.1469-1809.1936.tb02137.x.

[5] Friedlander, A. *Elementos de Programação Não-Linear*. Unicamp, 1994.

[6] Krulikovski, E. H. M. Análise teórica de máquinas de vetores suporte e aplicação a classificação de caracteres, Dissertação de Mestrado em Matemática, UFPR, 2017.

[7] Loh, W. Y., Tanner, M. A., Wolberg, W. H. Diagnostic Schemes for Fine Needle Aspirates of Breast Masses, *Analytical and Quantitative Cytology and Histology*, 10.3, p. 225–228, 1998.

[8] Luenberger, D. G. and Ye, Y. *Linear and Nonlinear Programming*. International Series in Operations Research & Management Science, Springer US, 2008.