Universidade Estadual de Campinas
Departamento de Estatística — IMECC
Disciplina: MI420 – Mineração de Dados
Professor: Dr. Ronaldo Dias
Acadêmicos: André Felipe B. Menezes (RA:229296)
Carolina da Silveira Bueno (RA:150282)

Uma comparação empírica entre *Support Vector Machines* e Regressão Logística

Campinas
Novembro de 2019

Resumo

Este trabalho apresenta uma discussão dos principais aspectos teóricos do *Support Vector Machine* (SVM), uma técnica de classificação proposta por pesquisadores da ciência da computação. O potencial do SVM foi explorado em um análise de dados reais, cujo objetivo é classificar genótipos de ratos com base nas suas medidas de proteínas. Os principais *Kernel* foram utilizados e o procedimento de validação cruzada foi empregado para estimação dos parâmetros de *tunning*. Os resultados foram comparados com o modelo de regressão logística e verificou-se que o SVM com os *Kernel* Polinomial e Radial apresentaram as melhores medidas de performance bem como uma excelente curva ROC.

Palavras chave: *Support Vector Machine*, regressão logística, validação cruzada.

Sumário

1 Introdução **3**

2 Metodologia **3**

3 Aplicação **5**

Referências **9**

1 Introdução

Uma tarefa muito usual na estatística é a formulação de modelos preditivos para classificar uma nova observação numa determinada categoria. Com o advento computacional várias abordagens tem sido propostas com intuito de obter modelos com alta acurácia.

Introduzido por Cortes e Vapnik (1995) o *Support Vector Machine* (SVM) é uma técnica matemática que tem sido bastante explorada na literatura em problemas de classificação. O SVM apresenta uma formulação puramente matemática diferenciando-se de outros modelos que fazem buscam mensurar a incerteza do fenômeno em estudo. De acordo com James et al. (2013) estudos empíricos mostram que o SVM tem apresentado uma performance bastante satisfatória em diferentes cenários.

Neste trabalho os principais aspectos da teoria que sustenta a formulação do SVM são apresentados. Além disso, ilustramos o potencial do SVM para um problema de classificação de genótipos de ratos. Os *Kernel* Linear, Polinomial, Radial e Sigmoide foram considerados e seus parâmetros de *tunning* foram estimados na amostra de treinamento utilizando validação cruzada. As principais medidas para avaliação do poder preditivo dos classificadores foram calculadas. Os resultados encontrados apontaram que o SVM com *Kernel* Polinomial e Radial apresentaram as melhores performance. Importante mencionar que toda a análise foi conduzida no ambiente R e a biblioteca `e1071` (MEYER, 2019) foi utilizada para ajuste do SVM.

2 Metodologia

Conforme discute James et al. (2013) o SVM é uma generalização do *Maximal Margin Classifier* e uma extensão do *Support Vector Classifier*.

Na sua essência, o SVM procura encontrar o melhor hiperplano que separa as observações nas duas classes. Este hiperplano é formado por uma projeção das variáveis preditoras em um espaço vetorial de maior dimensão, de modo que as classes sejam linearmente separáveis. Tal projeção é realizada utilizando funções *Kernel*.

Seja $\mathbf{X}$ uma matriz de dimensão $n \times p$ com n observações em um espaço p-dimensional, ou seja, uma matriz com os vetores $x_i = (x_{i1}, \ldots, x_{ip})^\top$, $i = 1, \ldots, n$. Considere que as n observações pertencem a duas classes, isto é, $y_1, \ldots, y_n \in \{-1, 1\}$. Suponha que exista um hiperplano separando as observações. Podemos expressar tal hiperplano da seguinte forma

$$y_i \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) > 0 \quad \forall i = 1 \ldots, n$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ são os parâmetros.

No *Maximal Margin Classifier* o hiperplano escolhido é aquele que apresenta maior

das menores distâncias entre as observações. Assim, os parâmetros deste hiperplano são obtidos resolvendo o seguinte problema de otimização

$$\begin{aligned} &\max_{\boldsymbol{\beta}, M} \\ &\text{sujeito à} \quad \sum_{j=1}^{p} \beta_j^2 = 1, \\ &y_i \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) > M \quad \forall i = 1 \ldots, n. \end{aligned}$$

Além de assumir que existe um hiperplano que separa perfeitamente as observações, esta formulação também é muito sensível a pequenas mudanças nos dados. Assim, surge o *Support Vector Classifier*, o qual apresenta um certa flexibilidade permitindo que algumas observações estejam do lado errado do hiperplano, porém dentro de uma margem, isto é, uma tolerância. Matematicamente, o problema de otimização passa a ser expresso por

$$\begin{aligned} &\max_{\boldsymbol{\beta}, \boldsymbol{\varepsilon}, M} \\ &\text{sujeito à} \quad \sum_{j=1}^{p} \beta_j^2 = 1, \\ &y_i \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\right) > M(1 - \varepsilon_i) \\ &\varepsilon_i \geq 0, \quad \sum_{i=1}^{n} \varepsilon_i \leq C \quad \forall i = 1 \ldots, n \end{aligned} \tag{1}$$

em que C é um parâmetro de *tunning* que controla o quão distante uma observação pode cair do lado errado do hiperplano. O caso onde as classes são separáveis, isto é, o *Maximal Margin Classifier* é obtido quando $C = \infty$.

Utilizando multiplicadores de Lagrange e as condições de Karush-Kuhn-Tucker é possível mostrar que o *Support Vector Classifier* possui uma representação envolvendo somente produto interno, De acordo com Hastie, Tibshirani e Friedman (2009) a nova representação é dada por

$$f(\boldsymbol{x}) = \beta_0 + \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$$

em que $\langle x, x_i \rangle = \sum_{j=1}^{p} x_j x_{ij}$.

Nesta representação têm-se n parâmetros $\alpha_i, i = 1, \ldots, n$ um para cada observação e para obter suas estimativas é necessário computar somente os $\binom{n}{2}$ produtos interno entre as observações. Portanto, o custo computacional é altamente barato.

Por outro lado, no SVM uma generalização do produto interno especificado por

uma função *Kernel* é utilizada. Ou seja,

$$f(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x, x_i)$$

em que $K(\cdot, \cdot)$ é a função *Kernel*.

A fórmula dos principais *Kernel* utilizados no SVM são apresentados na Tabela 1, sendo que $\boldsymbol{u} = (u_1, \ldots, u_n)$ e $\boldsymbol{v} = (v_1, \ldots, v_n)$. Observe que o caso do *Kernel* Linear têm-se o *Support Vector Classifier*. Note que os *Kernel* são controlados pelos parâmetros γ, d e c_0 que devem ser especificados pelo analista. Dessa forma, é usual utilizar validação cruzada na amostra de treinamento para determinar os melhores valores para os parâmetros.

Tabela 1: Principais *Kernel* do SVM.

Kernel	Fórmula	Parâmetros
Linear	$\boldsymbol{u}^\top \boldsymbol{v}$	—
Polinomial	$\left(\gamma \boldsymbol{u}^\top \boldsymbol{v} + c_0\right)^d$	γ, d, c_0
Radial	$\exp\left(\gamma \lvert \boldsymbol{u} - \boldsymbol{v} \rvert^2\right)$	γ
Sigmoide	$\tanh\left(\gamma \boldsymbol{u}^\top \boldsymbol{v} + c_0\right)$	γ, c_0

3 Aplicação

Para avaliar o desempenho empírico do SVM e comparar com a regressão logística um conjunto de dados relacionado ao genótipo de ratos foi analisado. O objetivo do estudo é desenvolver um modelo para classificar o genótipo de determinado rato com base nos seus valores de proteínas. A amostra consiste em 1073 observações e 69 níveis de proteínas (covariáveis), sendo que 52% dos ratos são do genótipo controle e 48% do genótipo trissômico (Ts65Dn). O conjunto de dados encontra-se disponível em <https://www.kaggle.com/ruslankl/mice-protein-expression/data>.

A Figura 1 ilustra o comportamento das variáveis explicativas utilizadas no modelo de regressão logística de acordo com o tipo de genótipo. É interessante perceber que algumas proteínas realizam uma boa discriminação em relação ao genótipo do rato. Por exemplo, os ratos do grupo Controle apresentam, em geral, menores valores da proteína APP do que os ratos do grupo Ts65Dn.

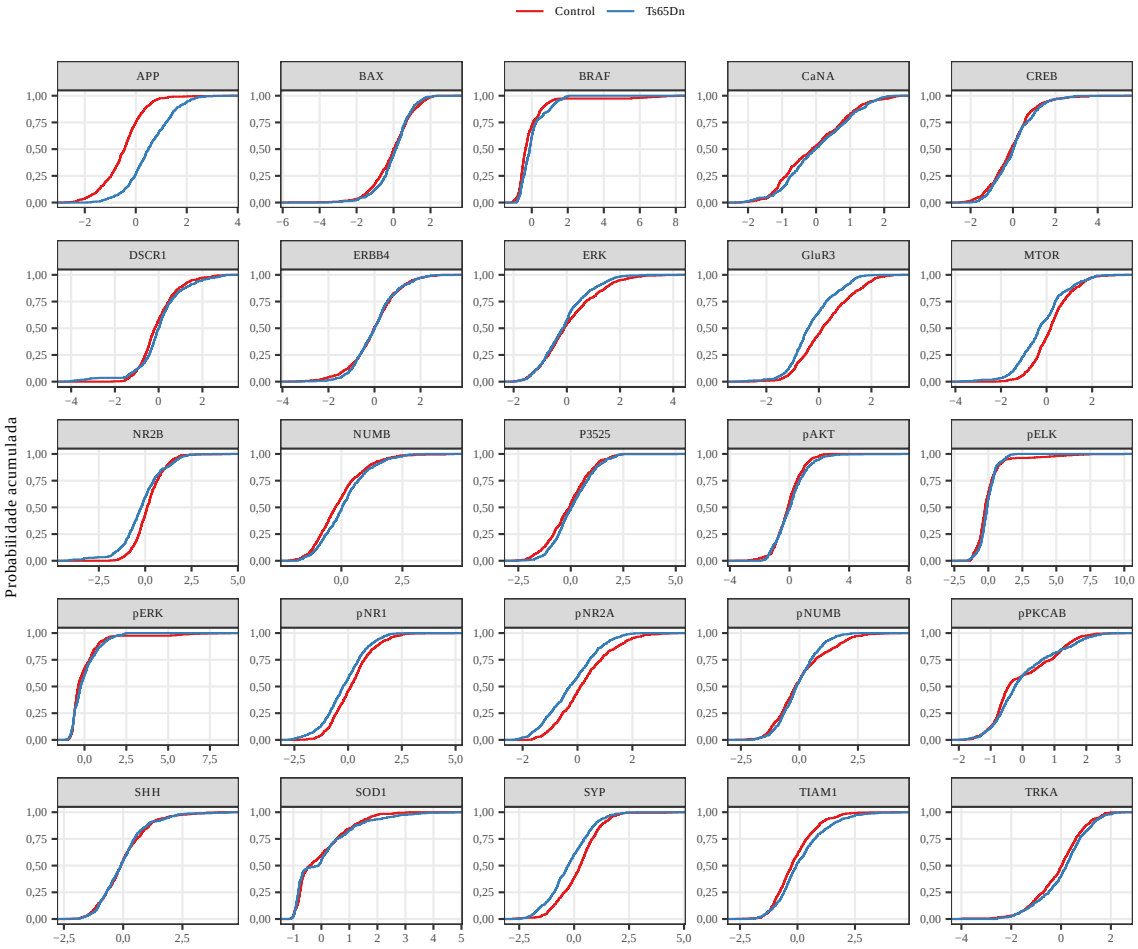

Figura 1: Comportamento de algumas variáveis explicativas conforme o tipo de genótipo.

Importante enfatizar que a análise foi conduzida considerando os seguintes aspectos:

- A amostra foi separada aleatoriamente em treino com 75% das observações e teste com 25% das observações restantes.
- No SVM os quatro *Kernel* apresentados na Tabela 1 foram considerados.
- Para estimação dos parâmetros do *Kernel* e do parâmetro de *tunning* C foi empregado validação cruzada com 10 *fold* na amostra de treinamento.
- Todas as covariáveis foram consideradas nos modelos SVM.
- Para o modelo de regressão logística a seleção das covariáveis foi realizada via *Backward-Stepwise*. As covariáveis selecionadas estão apresentada na 1.

A Tabela 2 apresenta os valores dos parâmetro de *tunning* considerados na validação cruzada (ver valores entre $\{\}$) e os melhores valores obtidos de acordo com o *Kernel*

do SVM. Por exemplo, entre todas as combinações dos valores dos parâmetros C, γ, c_0 e d a combinação $C = 0.5, \gamma = 0.01, c_0 = 1.0$ e $d = 4$ resultou no melhor ajuste, isto é, menor erro de predição estimado via validação cruzada na amostra de treinamento.

Tabela 2: Melhores valores para os parâmetros obtidos por validação cruzada, conforme o *Kernel* do SVM.

Parâmetro	Linear	Polinomial	Radial	Sigmoide
$C = \{0.1, 0.5, 1.0, 10\}$	0.5	0.5	10	10
$\gamma = \{0.01, 0.1, 1.0\}$	—	0.01	0.01	0.01
$c_0 = \{-1.0, 0.0, 1.0\}$	—	1.0	—	-1.0
$d = \{2, 3, 4, 5, 6\}$	—	4.0	—	—

Resultados das medidas de performance dos classificadores na amostra teste são apresentados na Tabela 3. Observa-se que a regressão logística apresentou pior desempenho preditivo, com um erro de predição estimado de $0,0712$. O SVM com os *Kernel* Radial e Polinomial apresentaram os melhores valores das medidas de performance considerada. O menor erro de predição estimado ($\mathrm{Err}_{\mathcal{T}}$) foi de $0,37\%$. Tais resultados corroboram com os resultados da Figura 2, onde as curvas ROC dos classificadores são comparadas.

Tabela 3: Medidas de performance dos classificadores na amostra teste.

Modelo	$\mathrm{Err}_{\mathcal{T}}$	AUC	Sens.	Espec.
Logística	0,0712	0,9780	0,9362	0,9206
Linear	0,0449	0,9853	0,9504	0,9603
Polinomial	**0,0037**	0,9944	**0,9929**	**1,0000**
Radial	**0,0037**	**0,9995**	**0,9929**	**1,0000**
Sigmoide	0,0337	0,9594	0,9787	0,9524

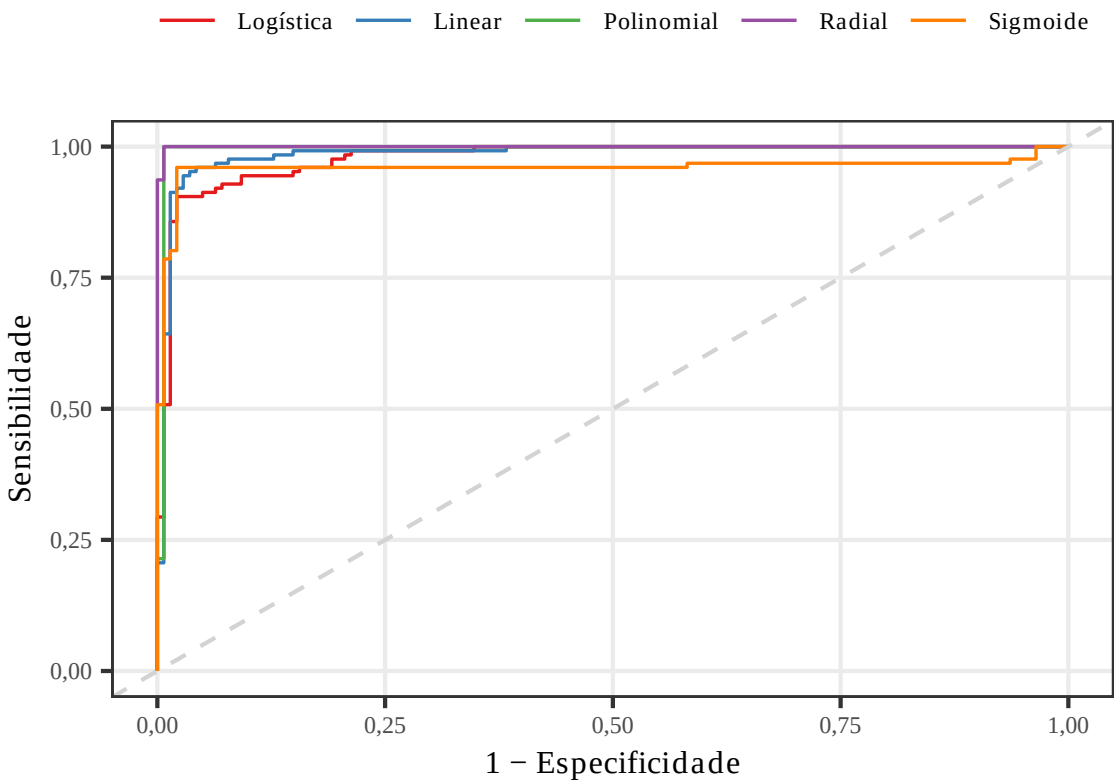

Figura 2: Curva ROC dos classificadores construída na amostra teste.

Conforme, os resultados apresentados e discutidos conclui-se que os métodos baseados em SVM apresentaram uma excelente performance e superaram o modelo de regressão logística em termos da capacidade preditiva. Entretanto, é importante enfatizar que o SVM não possui interpretabilidade que o modelo de regressão logística fornece.

Referências

CORTES, C.; VAPNIK, V. Support-vector network. **Machine Learning**, n. 20, p. 1–25, 1995.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining Inference and Prediction**. 2nd. ed. [S.l.]: Springer, 2009.

JAMES, G. et al. **An Introduction to Statistical Learning**. [S.l.]: Springer, 2013.

MEYER, D. **Support Vector Machines: The Interface to `libsvm` in package `e1071`**. [S.l.], 2019.