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)
Daniel Marques P. de Oliveira (RA:155083)

Classificadores Binários: Um estudo para classificação de pacientes com doença no coração

Campinas
Setembro de 2019

Resumo

Uma tarefa muito usual de um analista de dados é o desenvolvimento de modelos preditivos, em particular modelos para classificar uma nova observação em um determinada categoria. Devido o advento computacional, diferentes métodos de classificação foram propostos. No presente trabalho, foram discutidos aspectos teóricos de três classificadores binários – regressão logística, analise de discriminante linear e quadrática – e suas performances foram avaliadas em uma análise de dados reais, cujo objetivo é classificar pacientes com doença no coração. Métodos de validação cruzada e Bootstrap foram empregados para estimar o erro de predição esperado.

Palavras chave: regressão logística, análise discriminante, validação cruzada, Bootstrap.

Sumário

1 Introdução

Devido ao advento computacional, uma recente área na estatística, o Aprendizado Estatístico (*Statistical Learning*), ganhou maior notoriedade em diversas áreas do conhecimento. Combinando descobertas em paralelo com a ciência da computação, em particular o Aprendizado de Maquina (*Machine Learning*), os métodos dessa nova área são imprescindíveis para o analista de dados moderno.

De acordo com Hastie, Tibshirani e Friedman (2009), os problemas de aprendizagem podem ser grosseiramente categorizados em supervisionado e não supervisionado. No primeiro, o objetivo é predizer valores da variável resposta com base em um conjunto de covariáveis. Por outro lado, no aprendizado não supervisionado não há uma variável resposta e objetivo é descrever as associações e padrões entre um conjunto de variáveis.

O presente trabalho discute métodos de aprendizagem supervisionada. Em particular, é apresentada uma discussão teórica dos classificadores binários baseado na regressão logística e análise de discriminante. Também são discutidas técnicas para avaliar o desempenho dos classificadores. Finalmente, é mostrada uma aplicação detalhada dos métodos discutidos em um problema para classificação de pacientes com doença no coração.

2 Metodologia

2.1 Regressão Logística

Seja Y uma variável resposta dicotômica com dois níveis representados pelos valores 0 e 1. No modelo de regressão logistica, assume-se que $Y_i \mid \mathbf{x}_i \sim \text{Bernoulli}(\pi_i)$, em que o parâmetro $\pi_i = \Pr(Y_i = 1 \mid \mathbf{x}_i)$ é relacionado com os preditores $\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})^\top$ por meio da função de ligação *logit*. O modelo pode ser especificado por

$$Y_i \mid \mathbf{x}_i \sim \text{Bernoulli}(\pi_i), \quad \text{em que} \quad \pi_i = \frac{\exp\left(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}\right)}{1 + \exp\left(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}\right)} \tag{1}$$

em que i representa a característica do i-ésimo individuo, $i = 1, \ldots, n$ e $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ são os coeficientes de regressão associados as p covariáveis.

A estimação dos coeficientes de regressão, β_0 e $\boldsymbol{\beta}$, é realizada maximizando a função verossimilhança do modelo, a qual é definida por

$$\mathcal{L}(\beta_0, \boldsymbol{\beta} \mid \boldsymbol{y}) = \prod_{i=1}^{n} \pi_i^{y_i} \, (1 - \pi_i)^{1 - y_i} \, . \tag{2}$$

Infelizmente não existe uma solução analítica para obter as estimativas dos coefici-

entes, β_0 e $\boldsymbol{\beta}$, sendo assim necessário utilizar algoritmos numéricos para maximização da função (2) de forma a obtê-las.

Com as estimativas $\widehat{\beta}_0$ e $\widehat{\boldsymbol{\beta}}$ dos parâmetros é possível calcular as probabilidades preditas $\widehat{\pi}_i$ utilizando (1). Assim, a classificação do i-ésimo individuo é dada por

$$\widehat{y}_i = \begin{cases} 0, & \text{se } \widehat{\pi}_i < c \\ 1, & \text{se } \widehat{\pi}_i \geq c \end{cases}$$

em que $0 < c < 1$ é denominado ponto de corte. Usualmente considera-se $c = 0.5$, porém pode-se atribuir um valor para c conforme algum critério. Por exemplo, aquele que atinge a maior especificidade e sensibilidade.

2.2 Análise Discriminante Linear

Seja Y a variável resposta qualitativa com K classes distintas e seja $X = (X_1, \ldots, X_p)$ um vetor aleatório. Considere que π_k representa a probabilidade a priori de uma observação pertencer a k-ésima categoria de Y. Se $f_k(x) = \Pr(X = x \mid Y = k)$ denota a função densidade de probabilidade de X dado que $Y = k$, pelo teorema de Bayes temos que

$$p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k \, f_k(x)}{\sum_{i=1}^{K} \pi_i \, f_i(x)}. \tag{3}$$

Em geral, estimar π_k é trivial, com uma amostra aleatória da população Y uma estimativa natural para π_k é a proporção de observações na k-ésima classe. Por outro lado, estimar $f_k(x)$ tende a ser mais desafiador.

Conforme James et al. (2013) na análise discriminante linear (LDA) assume-se que a densidade $f_k(x)$ segue uma distribuição Normal multivariada $\mathrm{NM}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, em que μ_k é o vetor de médias da classe k e $\boldsymbol{\Sigma}$ é a matriz de covariância comum entre as K classes. Substituindo a densidade Normal multivariada em (3) e realizando algumas manipulações algébricas, pode-se mostrar que o classificador LDA atribui uma observação $X = x$ à classe para qual

$$\delta_k(x) = x^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k \tag{4}$$

é maior. Nota-se que $\delta_k(x)$ é uma função linear de x, este é o motivo do termo linear no nome deste método.

2.3 Análise Discriminante Quadrática

Ao contrário da LDA, a Análise Discriminante Quadrática (QDA) propõe que cada classe possua uma matriz de covariância diferente (MURPHY, 2012). Ou seja, assume-se que uma observação da k-ésima classe tenha distribuição normal multivariada com vetor de médias μ_k e matriz de covariância $\mathbf{\Sigma}_k$. Com isso, o classificador de Bayes (3) atribue uma observação x à uma classe para qual

$$\delta_k^*(x) = -\frac{1}{2}\log|\mathbf{\Sigma}_k| - \frac{1}{2}(x-\boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k (x-\boldsymbol{\mu}_k) + \log \pi_k \tag{5}$$

é maior. Nota-se que $\delta_k^*(x)$ é uma função quadrática de x, e não mais linear.

Conforme discute James et al. (2013), a escolha entre LDA e QDA vai depender da troca entre viés e variância. No cenário com p preditores, é necessário estimar $K(p+1)$ parâmetros no primeiro método, enquanto que, no segundo, precisa-se de $Kp(p+1)/2+k$. Consequentemente, LDA é muito menos flexível do que QDA, e portanto possui uma variância menor. Além disso, caso a suposição de homocedásticidade das K classes não seja válida, o método linear apresentará grande viés.

2.4 Comparação dos Classificadores

Apesar de possuirem motivações diferentes, regressão logística e LDA possuem um estreita relação (JAMES et al., 2013). Considere um problema com duas classes e um preditor. Sejam $p_1(x)$ e $1-p_1(x) = p_2(x)$ as probabilidades de $X = x$ pertencer à classe 1 e 2, respectivamente. Na LDA, o *logit* é dado por $\log\left(\frac{p_1(x)}{1-p_1(x)}\right) = \log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1 x$, em que c_0 e c_1 são funções de μ_1, μ_2 e σ^2. Por outro lado, na regressão logística (1) tem-se que $\log\left(\frac{p_1(x)}{p_2(x)}\right) = \beta_0 + \beta_1 x$.

Ambas estas equações são funções lineares de x, dessa forma produzem fronteiras de decisão lineares. A única diferença reside no fato de que os coeficientes da regressão logística são estimados por máxima verossimilhança, e na análise de discriminante as estimativas são obtidas estimando os parâmetros da distribuição normal. Essa relação também vale quando p é maior do que 1 (JAMES et al., 2013).

Como LDA e regressão logística diferem-se apenas na forma do ajuste, eles tendem a a apresentar resultados semelhantes. No entanto, isso nem sempre é verdade. LDA assume que as observações são retiradas de uma distribuição Normal, com variância comum entre as classes. Caso essa suposição seja verdadeira, ela pode exibir resultados melhores do que o método concorrente, e vice-versa. Comparando estes métodos com o KNN (*K-Nearest Neighbor*), é evidente que eles apresentam performance inferior quando a fronteira de decisão for extremamente não linear. Por outro lado, o KNN não mostra quais variáveis preditoras são importantes, dificultando seu uso para objetivos inferenciais.

Conforme James et al. (2013) ressaltam, a QDA possui maior flexibilidade do que os métodos linear (LDA e regressão logística), uma vez que assume que os limites de decisão são quadráticos.

3 Aplicação

Nesta seção é apresentada uma aplicação real e os classificadores discutidos nas seções anteriores são comparados. O conjunto de dados considerado refere-se a informações hospitalares e características pessoais de 303 pacientes, sendo 207 (68%) homens e 96 (32%) mulheres com idade entre 29 e 77 anos. Foram observadas 14 variáveis para cada paciente, sendo a característica de interesse (variável resposta) a presença de doença no coração. Um descrição detalhada das variáveis disponíveis é apresentada na Tabela A1. O conjunto de dados encontra-se disponível em <https://www.kaggle.com/ronitf/heart-disease-uci>.

Dos 303 pacientes 54% foram diagnosticados com doença no coração, ao passo que 46% não o foram. O objetivo desta análise é avaliar a performance dos métodos de regressão logística e análise de discriminante para classificar se determinado indivíduo apresenta doença no coração.

Conforme sugere Hastie, Tibshirani e Friedman (2009) separamos o conjunto em duas amostras aleatórias: amostra treino com 75% das observações e a amostra teste com 25% das observações restantes. A análise foi conduzida seguindo as etapas descritas a seguir:

1. **Seleção das Variáveis:** Com a amostra de treino, para cada modelo e método de classificação foi aplicada validação cruzada com $k = 10$ partições e calculadas várias medidas de performance.

2. **Performance dos Classificadores:** Utilizando a base de treino e com o "melhor" modelo para cada classificador foram utilizadas amostras Bootstrap para avaliar a acurácia das medidas de performance.

3. **Desempenho na Amostra Teste:** Por fim, com o modelo ajustado na amostra de treino o desempenho dos classificadores na amostra teste foi avaliado considerando várias medidas.

Para seleção do melhor subconjunto das variáveis preditoras foi conduzido um estudo de validação cruzada com $k = 10$ partições da amostra treino e calculadas algumas medidas para avaliar a performance dos classificadores. Este processo foi realizado para as $2^p - 1$ combinações de preditores (modelos). Importante destacar que nos modelos de regressão logística foram consideradas todas as $p = 13$ preditoras, resultando em 8.191 diferentes modelos distintos. Enquanto que nos classificadores

baseados na análise de discriminante tem-se apenas $p = 5$ preditores contínuos, logo um total de 31 modelos.

Na Tabela A2 estão reportadas as medidas de performance para os três melhores modelos conforme o classificador. Para cada classificador, o modelo com menor valor do erro de predição esperado $\left(\widehat{\text{Err}}\right)$ e maior AUC foi selecionado. Interessante notar que as variáveis `thalach` e `oldpeak` foram selecionadas nos três métodos de classificação.

Após a seleção das variáveis, avaliamos a acurácia das medidas de performance pelo método Bootstrap (EFRON; TIBSHIRANI, 1993). Foi utilizada $B = 2,000$ amostras Bootstrap da base de treino. Os resultados apresentados na Figura 1 apontam que o classificador com melhor desempenho é a regressão logística, a qual obteve menor valor mediano do erro de predição esperado (Err) e maiores valores medianos para AUC, sensibilidade e especificidade. Percebe-se também que a distribuição Bootstrap das medidas para os classificadores LDA e QDA são visualmente parecidas, com o LDA possuindo uma sutil vantagem.

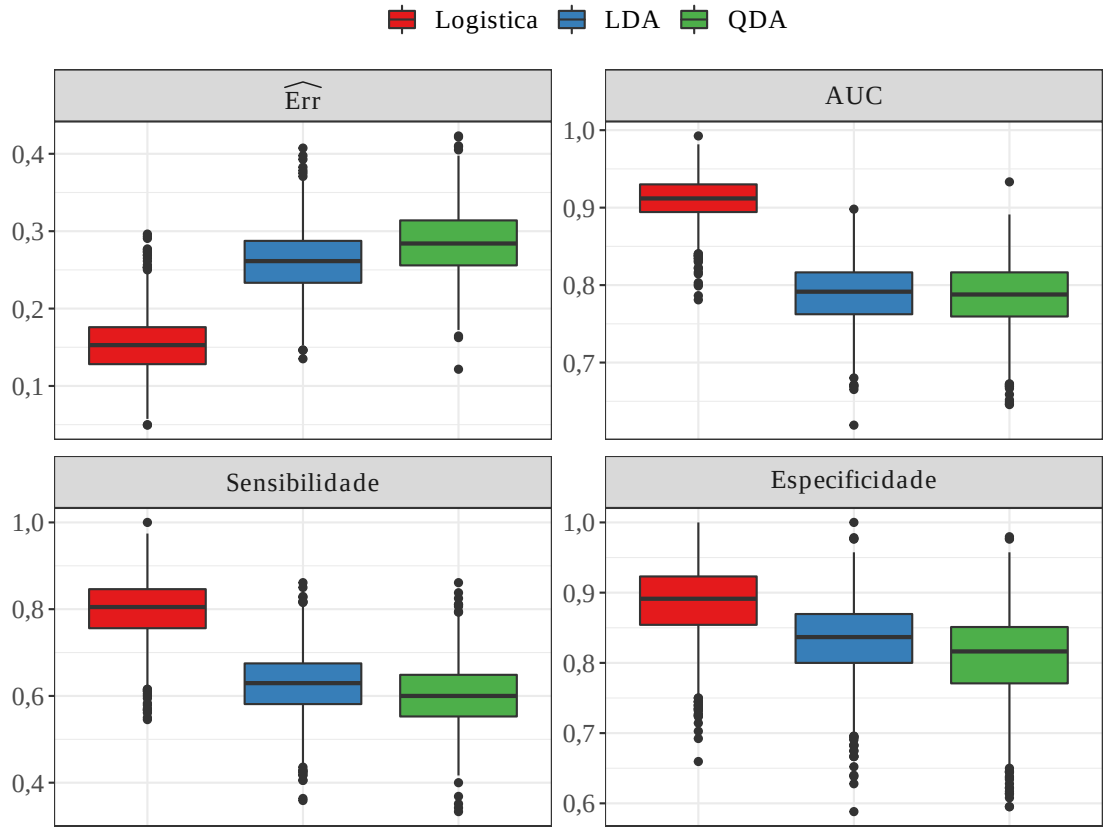

Figura 1: Distribuição Bootstrap das medidas de performance conforme classificador.

Finalmente, foi avaliado o desempenho dos classificadores na amostra teste. Na Figura 2 está apresentada a curva ROC dos classificados aplicados na amostra teste. Percebe-se que o classificador logístico se destaca dos classificadores LDA e QDA.

A Tabela 1 reporta as medidas de performance dos classificadores para a amostra teste. Observa-se que a regressão logística apresentou menor erro de predição ($\text{Err}_{\mathcal{T}}$), aproximadamente 0,13. Sob o ponto de vista prático este resultado significa que 13% dos individuos presentes na amostra teste foram classificados de forma errada em relação a presença ou não da doença no coração.

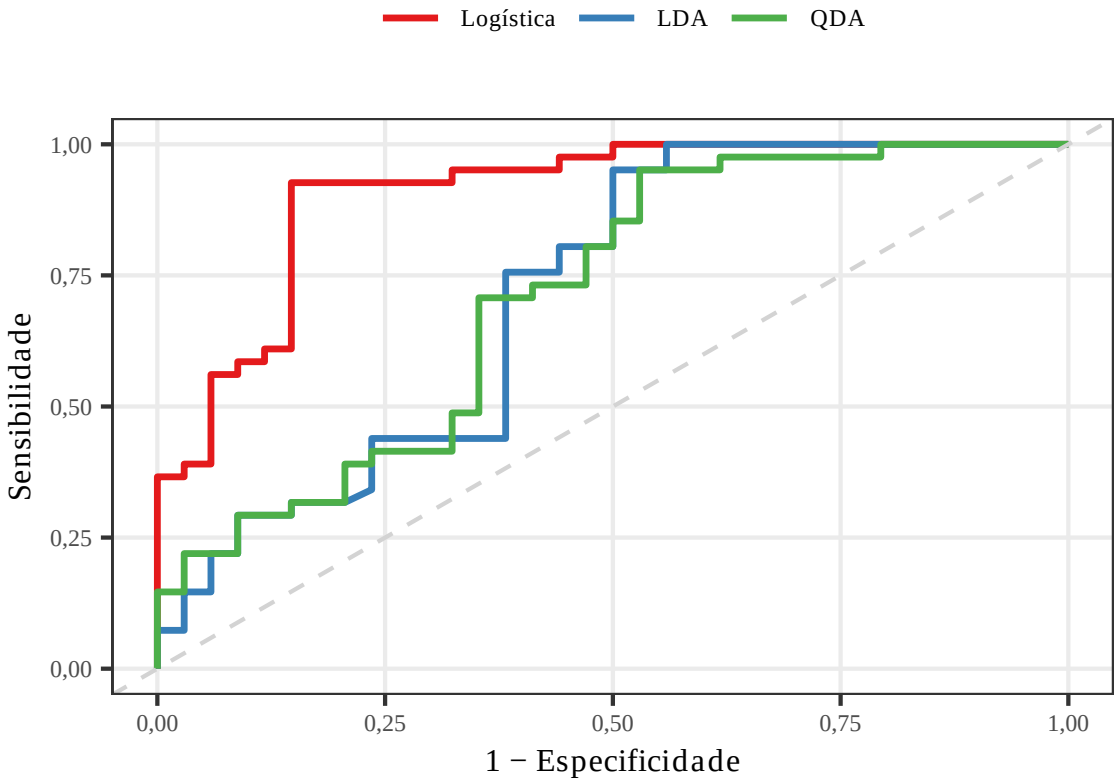

Figura 2: Curva ROC dos classificadores construída na amostra teste.

Tabela 1: Medidas de performance dos classificadores na amostra teste.

Modelo	$\text{Err}_{\mathcal{T}}$	AUC	Sens.	Espec.
Logística	0,1333	0,9067	0,8529	0,8780
LDA	0,3200	0,7120	0,5294	0,8049
QDA	0,3333	0,7109	0,6176	0,7073

Tendo em vista o estudo apresentado podemos concluir que para este conjunto de dados em específico, o método de classificação baseado na regressão logística demonstrou desempenho significativamente superior em relação aos métodos LDA e QDA, quando o objetivo é classificar indivíduos com doença no coração. Ressalta-se também que toda a análise foi realizada no software R e os códigos estão disponíveis em <https://github.com/AndrMenezes/dm2019>.

Referências

EFRON, B. Bootstrap methods: Another look at the jackknife. **The Annals of Statistics**, The Institute of Mathematical Statistics, v. 7, n. 1, p. 1–26, 01 1979.

EFRON, B.; TIBSHIRANI, R. **An Introduction to the Bootstrap**. Chapman & Hall/CRC, 1993.

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

JAMES, G. et al. **An Introduction to Statistical Learning**. Springer, 2013.

MURPHY, K. P. **Machine Learning: A Probabilistic Perspective**. MIT Press, 2012.

Apêndice

Tabela A1: Descrição das variáveis disponíveis.

Informação provida	**Variável**	**Tipo**
Doenção no coração (variável resposta)	`target`	(1 = Sim, 0 = Não)
Idade em anos	`age`	Numérica
Sexo	`sex`	Binária (1 = male, 0 = female)
Tipo de dor no peito	`cp`	Categórica (4 categorias)
Pressão arterial (mmHg)	`trestbps`	Numérica
Colesterol sérico (mg/dl)	`chol`	Númerica
Nível de açucar no sangue maior que 120 mg/dl	`fbs`	Binária (1 = sim, 0 = não)
Resultado eletrocardiográfico em repouso	`restecg`	Categórica (3 categorias)
Frequência cardiáca máxima	`thalach`	Numérica
Angina induzida pelo exercício	`exang`	Categórica (1 = Sim, 0 = Não)
Nível de depressão ST induzido por exercício em repouso	`oldpeak`	Numérica
Inclinação do segmento ST no pco do exercício	`slope`	Categórica (3 categorias)
Número de vasos principais	`ca`	Inteira
Talassemia	`thal`	Categórica (1 = Fixa, 2 = Reversível, 3 = Normal)

Tabela A2: Medidas de performance dos melhores modelos conforme o classificador.

Classificador	$\widehat{\text{Err}}$	$\overline{\text{err}}$	Sens.	Espec.	AUC	Preditores	p
Logística	0.1307	0.1213	0.8173	0.9128	0.9200	**`sex + cp + trestbps + fbs + restecg + thalach + oldpeak + ca + thal`**	9
	0.1307	0.1218	0.8173	0.9128	0.9185	`sex + cp + trestbps + restecg + thalach + oldpeak + ca + thal`	8
	0.1311	0.1296	0.8164	0.9128	0.9115	`sex + cp + trestbps + thalach + ca + thal`	6
LDA	0.2473	0.2486	0.6318	0.8538	0.7918	**`thalach + oldpeak`**	2
	0.2523	0.2500	0.6309	0.8455	0.7906	`age + thalach + oldpeak`	3
	0.2644	0.2583	0.6318	0.8224	0.7881	`chol + thalach + oldpeak`	3
QDA	0.2723	0.2758	0.6145	0.8224	0.7982	**`trestbps + thalach + oldpeak`**	3
	0.2777	0.2783	0.5755	0.8455	0.7869	`thalach + oldpeak`	2
	0.2814	0.2705	0.6145	0.8058	0.7970	`age + trestbps + thalach + oldpeak`	4

$\widehat{\text{Err}}$: estimativa do erro de predição esperado (*expected test error*).
$\overline{\text{err}}$: estimativa do erro de predição de treinamento (*training error*).
p: número de preditores.

Avaliação da Performance do Classificador

Na amostra de treinamento, seguindo a recomendação de Hastie, Tibshirani e Friedman (2009) o erro de predição esperado (Err) e o erro de predição de treinamento ($\overline{\text{err}}$), para cada classificador, foram estimados utilizando o método Bootstrap proposto por Efron (1979). Além disso, a área sob a curva ROC (AUC), sensibilidade e especificidade também foram estimadas.

Considere a seguinte matriz de confusão construída a partir da amostra de teste.

Observado	**Predito**	
	0	1
0	A	C
1	B	D

Então, as medidas mencionadas são calculadas para a amostra de teste da seguinte forma

$$\text{Err} = \frac{C+B}{A+B+C+D}, \quad \text{Sens.} = \frac{A}{A+C}, \quad \text{Espec.} = \frac{A}{A+C} \tag{6}$$

enquanto que a AUC foi calculada utilizando o método de integração por trapézios.