Aprendizado de Máquina

UFMG EST171 - 2ª Lista de exercícios

Eduardo Elias Ribeiro Junior

04 de outubro de 2016

Exercício 1

> *Baixe o conjunto de dados `titanic.txt`. Cada observação deste banco é relativa a um passageiro do Titanic. As covariáveis indicam características deste passageiros; a variável resposta indica se o passageiro sobreviveu ou não ao naufrágio.*
> *Seu objetivo é criar classificadores para predizer a variável resposta com base nas covariáveis disponíveis. Para tanto, você deverá implementar os seguintes classificadores, assim como estimar seus riscos via conjunto de teste:*

O conjunto de dados `Titanic` é apresentado na Tabela 1. As características dos passageiros disponíveis nesse conjunto são: classe econômica - **`Class`**, de quatro categorias; sexo - **`Sex`**, de duas categorias; e idade do passageiro - **`Age`**, categorizada em adultos ou crianças. Para todos os cruzamentos dessas covariáveis têm-se a frequência de sobreviventes e não sobreviventes - **`Survived`**.

Tabela 1: Tabela de frequência dos passageiros do Titanic

			Survived	No	Yes	**Total**
Class	**Sex**	**Age**				
1st	Female	Adult		4	140	144
		Child		0	1	1
	Male	Adult		118	57	175
		Child		0	5	5
2nd	Female	Adult		13	80	93
		Child		0	13	13
	Male	Adult		154	14	168
		Child		0	11	11
3rd	Female	Adult		89	76	165
		Child		17	14	31
	Male	Adult		387	75	462
		Child		35	13	48
Crew	Female	Adult		3	20	23
		Child		0	0	0
	Male	Adult		670	192	862
		Child		0	0	0

A análise gráfica descritiva do conjunto de dados é realizada na **??** onde exibe-se, acima, as frequências das categorias de cada variável presente no conjunto de dados e, abaixo, as proporções da tabela de contingência em cada combinação das variáveis dispostas em áreas retangulares. Primeiramente observa-se que a conjunto de dados não é balanceado em praticamente nenhuma variável, esse desbalanço é maais notável para a idade dos passageiros, onde observa-se que, aproximadamente, `paste0(round(prop.table(table(dados$Age))[1], 3)*100, "%")` são passageiros adultos. Já no gráfico abaixo, nota-se que praticamente todas as mulheres da primeira classe sobreviveram e que houveram menos passageiros sobreviventes do sexo masculino. Prelimi-

narmente, pode-se imaginar que todas as covariáveis observadas, com exceção de **Age**, podem ser utéis para a classificar se um passageiro é sobrivente ou não, ou ainda, calcular sua probabilidade de sobrevida em uma possível tragédia como essa.

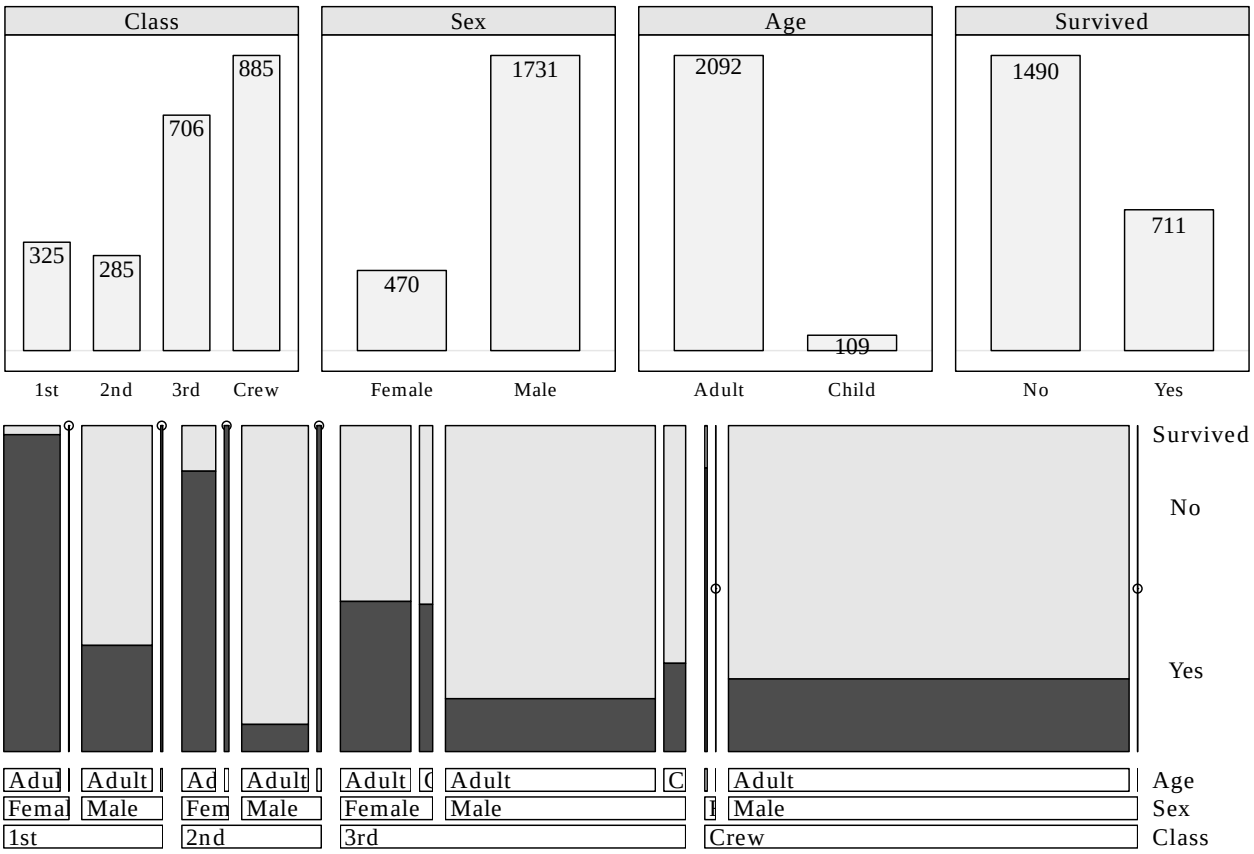

Figura 1: Gráficos descritivos da base de dados. Frequências observadas em cada variável de Titanic (superior) e Representação da tabela de contingência de forma hierárquica (inferior).

Para dar sequência a obtenção de classificadores, será realiza a partição da base de dados em dois conjuntos. Um para ajuste do classificador e outro para validação deste. A partição será realizada a partir da função implementada para tal finalidade, os detalhes da implementação dessa função são exibidos no complemento online do trabalho.

```
## Particionando o conjunto de dados
dasplit <- mysplit(dados, percent = c(0.7, 0.3), seed = 1994)

## Número de observações em cada partição
sapply(dasplit, nrow)
```

```
## [1] 1541  660
```

```
## Atribuindo as partições em objetos de nome sujestivo
da.train <- dasplit[[1]]
da.teste <- dasplit[[2]]
```

```
## Tranformando para inteiro, para utilização de alguns métodos
X.train <- sapply(da.train, as.integer)
X.teste <- sapply(da.teste, as.integer)
```

Regressão Logística

O primeiro classificador a ser construído, será fundamentado sob a teoria dos modelos lineares generalizados. Associadoremos à variável resposta (`Survived`), condicional ao vetor de covariáveis (`Class`, `Sex` e `Age`), a distribuição Binomial de parâmetro π, onde π é função não linear (inversa da função logística) dos efeitos das covariáveis. A especificação do modelo é descrita abaixo.

$$\text{Survived}_i \mid \text{Class}_i, \text{Sex}_i, \text{Age}_i \sim \text{Binomial}(\pi_i)$$
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_{11}X_{2st,i} + \beta_{12}X_{3st,i} + \beta_{13}X_{Crew,i} + \beta_2 X_{Male,i} + \beta_3 X_{Child,i}$$

em que i representa as características do i-ésimo indivíduo, $i = 1, 2, \ldots, n$. $\underline{\beta}$ é o vetor dos parâmetros que representam os efeitos das covariáveis. E $X_{j,i}$ é uma variável binária que assume, para o i-ésimo indivíduo: 1 se a variável `Class` é igual a j e 0 caso contrário, para $j = 2St$, $3St$ e $Crew$; 1 se a variável `Sex` é igual a $Male$ e 0 caso contrário, para $j = Male$; e 1 se a variável `Age` é igual a $Child$ e 0 caso contrário, para $j = Child$.

Ajustando esse modelo ao conjunto de treinamento, `da.train` têm-se os seguintes coeficientes estimados, com seu erro padrão calculado a partir da aproximação quadrática da versossimilhança e nível de significância do teste de Wald:

Tabela 2: Coeficientes estimados e teste de Wald para o modelo Logístico

	Estimate	Std. Error	z value	Pr(>|z|)
β_0	2.15	0.2057	10.45	1.43e-25
β_{11}	-0.9756	0.2381	-4.098	4.175e-05
β_{12}	-1.819	0.2102	-8.652	5.067e-18
β_{12}	-0.7844	0.1891	-4.147	3.37e-05
β_2	-2.57	0.1733	-14.83	9.324e-50
β_3	1.053	0.2892	3.641	0.0002711

Note que o modelo logístico conforme descrito não é essencialmente um classificador, pois é um modelo para a probabilidade. Utilizando dessa probabilidade predita pelo modelo logística para classificação fez-se a classificação da forma $\hat{\pi}_i > 0.5$ classifica-se como sobrevivente (`Survived` = `Yes`) e não sobrevivente (`Survived` = `No`) caso contrário. Com essa regra de classificação obtêm-se a proporção de 0.2348485 classificações incorretas no conjunto de teste.

Regressão Linear

Similarmente à regressão logística este também é um modelo fundamentado na teoria dos modelos lineares generalizados, porém é definido no plano cartesiano, por assumir a distribuição Normal à variável de interesse condicionada as covariável, e sendo assim tem solução geométrica análitica (de mínimos quadrados). A regressão Gaussiana é o único modelo dessa classe com essa característica. O modelo é definido conforme especificação abaixo:

$$\text{Survived}_i \mid \text{Class}_i, \text{Sex}_i, \text{Age}_i \sim \text{Normal}(\mu_i, \sigma^2)$$
$$\mu_i = \beta_0 + \beta_{11}X_{2st,i} + \beta_{12}X_{3st,i} + \beta_{13}X_{Crew,i} + \beta_2 X_{Male,i} + \beta_3 X_{Child,i}$$

em que i representa as características do i-ésimo indivíduo, $i = 1, 2, \ldots, n$. $\underline{\beta}$ é o vetor dos parâmetros que representam os efeitos das covariáveis. E $X_{j,i}$ é uma variável binária que assume, para o i-ésimo indivíduo: 1 se a variável `Class` é igual a j e 0 caso contrário, para $j = 2St$, $3St$ e $Crew$; 1 se a variável `Sex` é igual a $Male$ e 0 caso contrário, para $j = Male$; e 1 se a variável `Age` é igual a $Child$ e 0 caso contrário, para $j = Child$.

Note que claramente esse não seria um modelo adequado uma vez que o domínio da distribuição Normal são os reais e, sendo assim, pode haver predições negativas e maiores que 1 para a média μ_i. Isso é contemplado no modelo generalizado logístico, porém quando o interesse é somente predição, ambos são classificadores que devem ser avaliado, mesmo que a regressão linear tenha características inadequadas ao conjunto de dados.

Com o modelo ajustado ao conjunto de dados de treino, exibe-se na Tabela 3 os coeficientes estimados juntamente com seu erro padrão e respectivo teste de Wald.

Tabela 3: Coeficientes estimados e teste de Wald para o modelo Gaussiano

	Estimate	Std. Error	t value	Pr(>|t|)
β_0	1.906	0.0306	62.28	0
β_{11}	-0.1728	0.03918	-4.41	1.103e-05
β_{12}	-0.3034	0.03309	-9.17	1.476e-19
β_{12}	-0.1592	0.03319	-4.797	1.769e-06
β_2	-0.5149	0.02755	-18.69	2.142e-70
β_3	0.1773	0.04749	3.734	0.0001956

Novamente, assim como realizado no modelo logístico, faremos a classificação a partir da probabilidade predita pelo modelo Gaussiano. Conforme discutido anterior essa predição de probabilidade tem diversas falhas, sendo a mais agravante não respeitar o espaço paramétrico (pode-se ter probabilidade maiores que 1 e menores que 0), porém utilizando a regra da classificação $\hat{\mu_i} > 0.5$ classifica-se como sobrevivente (`Survived` = `Yes`) e não sobrevivente (`Survived` = `No`) caso contrário, temos uma proporção de classificações incorretas no conjunto de teste de 0.2348485.

Naive Bayes

Este é um classificador fundamentado a partir do teorema de bayes. Como o principal objetivo em classificação é estimar $\Pr[Y = c \mid X] \, \forall \, c \in \mathbb{C}$, sendo $\mathbb{C}$ o conjunto de categorias da variável resposta, utilizando o Teorema de Bayes temos:

$$\Pr(Y = c \mid x) = \frac{f(\underline{x} \mid Y = c) \Pr(Y = c)}{\sum_{s \in \mathbb{C}} f(\underline{x} \mid Y = s) \Pr(Y = s)}$$

e assim calcula-se $\Pr(Y = c \mid x)$ estimando $Pr(Y = c)$, comumente como proporções amostrais e $f(\underline{x} \mid Y = c)$, onde o classificador Naive Bayes supõe independência condicional

$$f(\underline{x} \mid Y = c) = \prod_{j=1}^{p} f(x_j \mid Y = y)$$

Para os dados do Titani todas as covariáveis são categóricas, portanto assume-se $f(x_j \mid Y = c)$ como uma distribuição Multinomial (ou Binomial no caso de duas categorias), assim têm-se 16 probabilidades a serem calculadas conforme exibe-se na Tabela 4.

Tabela 4: Probabilidades estimadas para cada categoria de cada covariável condicional a Survived

	1st	2nd	3rd	Crew	Female	Male	Adult	Child
No	0.0822	0.107	0.346	0.464	0.0783	0.922	0.964	0.0358
Yes	0.284	0.174	0.237	0.306	0.493	0.507	0.913	0.0868

Com a regra da classificação $\hat{\Pr}(\texttt{Survived=Yes}|\ \underline{x}_i^t) > 0.5$ classificado como sobrevivente (`Survived` = `Yes`) e não sobrevivente (`Survived` = `No`) caso contrário, temos uma proporção de classificações incorretas no conjunto de teste de 0.230303.

Análise Discriminante Linear

Em análise discriminante linear de Fisher, ainda utiliza-se o teorema de Bayes da mesma forma como descrito na seção Naive Bayes, porém assume-se para $f(\underline{x} \mid Y = c)$ a distribuição Normal multivariada de parâmetros $\underline{\mu_c}$ e matriz de variâncias e covariâncias Σ comum para a toda categoria $c \in \mathbb{C}$

Perceba-se que aparentemente essa abordagem não parece satisfatória uma vez que todo o conjunto de covariáveis é categórica, assim estamos assumindo uma distribuição Normal para algo que é claramente discreto. Todavia como já discutido na seção Regressão Linear o interesse é apenas preditivo, e sendo assim, podemos construir uma regra de classificação que será posteriormente avaliada. O classificador ajustado ao conjunto de treinamento é exibido abaixo.

```
## Call:
## lda(Survived ~ ., data = da.train)
##
## Prior probabilities of groups:
##        No       Yes
## 0.6709929 0.3290071
##
## Group means:
##      Class2nd  Class3rd ClassCrew   SexMale   AgeChild
## No  0.1073501 0.3462282 0.4642166 0.9216634 0.03578337
## Yes 0.1735700 0.2366864 0.3057199 0.5069034 0.08678501
##
## Coefficients of linear discriminants:
##                  LD1
## Class2nd  -0.8283529
## Class3rd  -1.4542324
## ClassCrew -0.7630442
## SexMale   -2.4677185
## AgeChild   0.8498164
```

Com a regra da classificação $\hat{\Pr}(\texttt{Survived=Yes}|\ \underline{x}_i^t) > 0.5$ classificado como sobrevivente (`Survived` = `Yes`) e não sobrevivente (`Survived` = `No`) caso contrário, temos uma proporção de classificações incorretas no conjunto de teste de 0.2348485.

Análise Discriminante Quadrática

A análise discriminante quadrática de Fisher segue o mesmo princípio da análise discriminante linear, porém flexibiliza $f(\underline{x} \mid Y = c)$ estimando uma matriz de variâncias e covariâncias *Sigma* para cada classe $c \in \mathbb{C}$, ou

seja, nessa abordagem assume-se que

$$[\underline{x} \mid Y = c] \sim \text{Normal}(\mu_c, \Sigma_c) \quad \forall\, c \in \mathbb{C}$$

Novamente as mesmas considerações feitas na abordagem via análise discriminante linear se aplicam. O resultado do classificador ajustado aos dados de treino é exibido abaixo

```
## Call:
## qda(Survived ~ ., data = da.train)
##
## Prior probabilities of groups:
##        No       Yes
## 0.6709929 0.3290071
##
## Group means:
##      Class2nd  Class3rd ClassCrew   SexMale   AgeChild
## No  0.1073501 0.3462282 0.4642166 0.9216634 0.03578337
## Yes 0.1735700 0.2366864 0.3057199 0.5069034 0.08678501
```

Analógo com o classificador construído a partir da análise discriminante linear. Utilizamos a regra da classificação $\hat{\Pr}(\texttt{Survived=Yes} \mid \underline{x}_i^t) > 0.5$ classificado como sobrevivente (`Survived` = `Yes`) e não sobrevivente (`Survived` = `No`) caso contrário, temos uma proporção de classificações incorretas no conjunto de teste de 0.2681818.

K-NN: k-Nearest Neighbor

Este método se diferencia dos demais por ser totalmente não paramétrico, não há suposição de distribuição ou quaisquer parâmetros a ser estimados. O método se baseia em classificar ou predizer os valores da variável de interesse a partir dos valores de seus `k` vizinhos, no caso de classificação, classifica-se uma nova observação como a moda das `k` observações mais próximos e no caso de predição (considerar uma variável resposta não categórica) prediz-se com base na média. A proximidade é dada por distâncias euclidianas e o valor de quantas observações devem ser usadas para classificação ou predição, `k`, é realizada separando 20% da base de treino para validação.

Nesse trabalho utiliza-se ainda, o algoritmo `ANN` (*Approximate Nearest Neighbor Searching*), proposto em 1993 por Arya S. and Mount D.M. para aplicação do KNN em problemas de alta dimensão onde os algoritmos convencionais para cálculo de distância são ineficientes. Esse algoritmo é implementado em `C++` e está disponível em R no pacote `FNN` (via argumento `algorithm = "kd_tree"`, nas funções do pacote)[1]. Utilizou-se deste algoritmo para escolhe do melhor `k` via validação cruzada.

[1] *Paper:* http://www.cs.umd.edu/~mount/Papers/soda93-ann.pdf
Page: https://www.cs.umd.edu/~mount/ANN/.

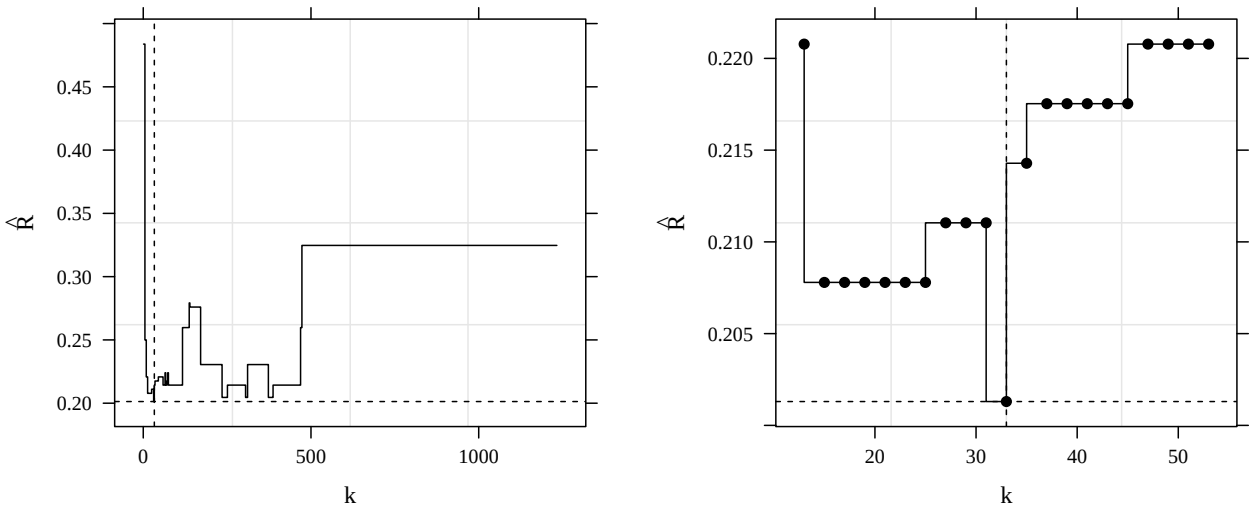

Figura 2: Proporção de classificações incorretas pelo classificador KNN com diferentes k's. Todos os possíveis k's ímpares (esquerda) e apenas os k's próximos do k ótimo.

Na Figure 2 utilizamos o método considerando diferente número de vizinhos, `k`, para determinar o `k` definitivo a ser utilizado para classificação. Na figura são exibidos a proporção de classificações incorretas denotada por $\hat{R}$ no eixo das ordenadas. O número de vizinhos que proporcionou a menor proporção de classificações incorretas foi de 33 vizinhos, com uma proporção de 0.2287879.

Comparação dos métodos

Para comparação dos métodos será utilizado, além da proporção de classificações incorretas as medidas de sensibilidade e especificadade calculadas a partir da matriz de confusão obtida por cada classificador.

Tabela 5: Comparação dos métodos utilizando o ponto de corte usual de 0.5

	GLM	LM	NB	LDA	QDA	KNN
Prop. de Acertos	0.7652	0.7652	0.7697	0.7652	0.7318	0.7712
Sensibilidade	0.4608	0.4608	0.4755	0.4608	0.5882	0.2941
Especificidade	0.9013	0.9013	0.9013	0.9013	0.7961	0.9846
Escore	0.7231	0.7231	0.729	0.7231	0.712	0.7053

Na Tabela 5 são exibidas as medidas para comparação dos classificadores com base na probabilidade de corte de 0.5. Todos os métodos apresentados neste trabalho contém alguma medida que pode ser interpretada como uma estimativa da probabilidade. O modelo logístico é naturalmente um modelo para prever a probabilidade, o Gaussiano ainda que não respeite o espaço paramétrico da probabilidade a estima, o Naive Bayes assim como as análises discriminantes de Fisher utilizam do teorema de Bayes para estimar probabilidades e no K-NN podemos estimar essa probabilidade como a proporção dos k vizinos mais próximos em certa categoria.

As avaliações até agora, foram realizadas com o classificador obtidos considerando o ponto de corte 0.5, nas probabilidades estimadas. Todavia pode-se encontrar o ponto de corte ótimo para cada método considerando um conjunto de validação. Ilustramos os resultados desse procedimento a seguir.

Na Figure 3 apresentam-se as curvas ROC para todos os classificadores, essas foram contruídas a partir dos classificadores ajustados ao conjunto de treino, removendo 20% dele para validação cruzada, ou seja,

os resultados exibidos são com base na classificação de 20% do conjunto de treinamento (308 observações). A partir deste da avaliação da curva ROC indentificou-se os pontos de corte que maximizam a soma entre sensibilidade e especificidade do classificador. Além disso pode-se notar também as semelhanças e características deles.

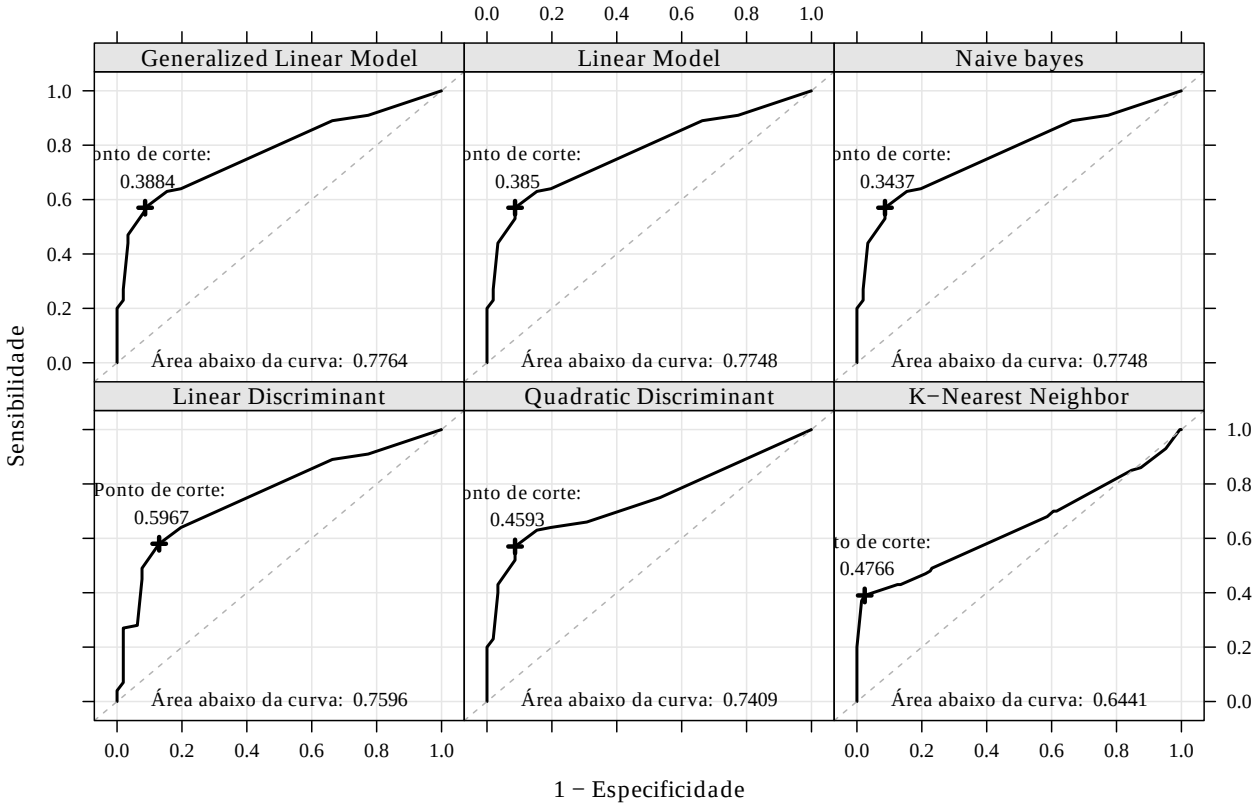

Figura 3: Curvas ROC (Receiver Operating Characteristic) para cada modelo de classificação com indicação do melhor ponto de corte e respectivo AUC (Area Under Curve) para o conjunto de validação.

Contruindo a regra de classificação a partir dos pontos de cortes ótimos exibidos na Figure 3 temos na Tabela 6 os resumos da matriz de confusão, quando classificadas as 660 observações do conjunto de teste.

Tabela 6: Comparação dos métodos via resumos da matriz de confusão da classificação vs. teste, utilizando o ponto de corte ótimo (obtido por validação cruzada)

	GLM	LM	NB	LDA	QDA	KNN
Prop. de Acertos	0.7439	0.7439	0.7439	0.7652	0.7318	0.7712
Sensibilidade	0.5686	0.5686	0.5686	0.4608	0.5882	0.2941
Especificidade	0.8224	0.8224	0.8224	0.9013	0.7961	0.9846

Note que embora na validação cruzada o método KNN tenha sido o de pior desempenho, quando ajustado a todo conjunto de treinamento e utilizado para classificação do conjunto de teste essa abordagem obteve os melhores resultados. Todavia cabe salientar que o método KNN é o menos parcimonioso de todos os avaliados, pois sua especificidade é muito alta em contraste com sua sensibilidade que é muito baixa, em comparação com os demais. Outro fato interessante dessa análise é que os classificadores baseados no modelo Logístico, Linear e Naive Bayes resultaram nas mesmas classificações e, além disso os demais

resultados, com excessão do KNN, também foram bastante similares, isso sugere que a cojunto de dados analisado apresenta um comportamento das covariáveis com dentro das categorias bastante característico, o que leva os classificadores a predizer da mesma forma.

Material suplementar

Todos os códigos (para manipulação, ajustes e gráficos) exibidos neste trabalho estão disponíveis no endereço https://jreduardo.github.io/est171-ml/.