MAE 5905: Introdução à Ciência de Dados

Pedro A. Morettin

Instituto de Matemática e Estatística
Universidade de São Paulo
pam@ime.usp.br
http://www.ime.usp.br/∼ pam

Aula 7

25 de março de 2026

Sumário

1 Exemplo-Regularização

2 VC e Bootstrap

3 Classificação Clássica

Regularização - EMQ

- **Exemplo.** Vamos considerar o conjunto de dados **esforco**, centrando o interesse na predição da variável resposta Y: VO2 (consumo de oxigênio em ml/(kg.min)) com base nas variáveis preditoras X_1: Idade (em anos), X_2: Peso (em kg), X_3: Superfície corpórea e X_4: IMC (índice de massa corpórea em kg/m^2) ($n = 126$)
- Ajustando o modelo via mínimos quadrados ordinários, obtemos os coeficientes:

Coefficients	Estimate	Std. Error	t value	P-value
Intercept	14.92204	6.69380	2.229	0.0276 *
Idade	-0.01005	0.02078	-0.483	0.6297
Peso	0.05233	0.11760	0.445	0.6571
Sup.Corp.	-1.40678	6.25353	-0.225	0.8224
IMC	-0.20030	0.16104	-1.244	0.2160

Residual standard error: 2.822 on 121 degrees of freedom

Multiple R-squared: 0.03471, Adjusted R-squared: 0.002799

Regularização - Ridge

- Os coeficientes correspondentes obtidos por meio de regularização **Ridge** são

Intercept	5.185964640
Idade	-0.000133776
Peso	-0.006946405
Sup.Corp	-0.295094364
IMC	-0.022923850

- O valor do coeficiente de regularização $\lambda = 0,82065$, mostra que as estimativas para os coeficientes de Idade e Peso foram encolhidas para zero, enquanto aquelas correspondentes à Sup.Corp tem peso maior do que as demais.
- Neste caso, a raiz quadrada do **erro quadrático médio** (*root mean squared error*) e o **coeficiente de determinação** são, respectivamente $RMSE = 0,928$ e $R^2 = 0,235$.

Regularização Lasso

- Os coeficientes correspondentes obtidos por meio de regularização Lasso são

Intercept	4.95828012
Idade	.
Peso	-0.01230145
Sup.Corp	.
IMC	-0.02011871

- O valor do coeficiente de regularização $\lambda = 0,0257$ mostra que as estimativas dos coeficientes Idade e Sup. Corp foram efetivamente encolhidas para zero.
- Neste caso RMSE e R^2 são, respectivamente, $0,927$ e $0,228$.

Regularização Elastic Net

- Os coeficientes correspondentes obtidos por meio de regularização Elastic Net são:

Intercept	4.985532570
Idade	.
Peso	-0.009099925
Sup.Corp	-0.097034844
IMC	-0.023302254

- Os parâmetros de suavização estimados foram $\alpha = 0,1$ e $\lambda = 0,227$ e também indicam que o coeficiente associado à Idade foi encolhido para zero.
- Também obtemos RMSE=0,927 e $R^2 = 0,228$ neste caso.
- Os três métodos de regularização têm desempenhos similares quando vistos pelas óticas do RMSE e do R^2.

Validação Cruzada

- Validação cruzada é a denominação atribuída a um conjunto de técnicas utilizadas para avaliar o erro de previsão de modelos estatísticos. O erro de previsão é uma medida da precisão com que um modelo pode ser usado para prever o valor de uma nova observação, ou seja diferente daquelas utilizados para o ajuste do modelo.
- Em modelos de regressão o erro de previsão é definido como

$$EP = E(y - \widehat{y})^2,$$

em que y representa uma nova observação e $\widehat{y}$ é a previsão obtida pelo modelo.

- erro quadrático médio dos resíduos pode ser usado como uma estimativa do erro de previsão (EP), mas tende, em geral, a ser muito otimista, ou seja, a subestimar o seu verdadeiro valor. Uma razão é que os mesmos dados são utilizados para ajustar e avaliar o modelo.
- No processo de validação cruzada, o modelo é ajustado a um subconjunto dos dados (**subconjunto de treinamento**) e o resultado é empregado num outro subconjunto (**subconjunto de teste**) para avaliar se ele tem um bom desempenho ou não.

Validação Cruzada

Um algoritmo utilizado nesse processo é o seguinte (Efron e Tibshirani, 1993):

1) Dadas n observações, $y_1, \ldots, y_n$, o modelo é ajustado n vezes, em cada uma delas eliminando uma observação e o valor previsto, denotado por $\widehat{y}_{-i}$, para essa observação, é calculado com base no resultados obtido com as demais $n-1$.
2) O erro de previsão é estimado por

$$VC_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_{-i})^2. \quad (1)$$

Esse tipo de validação cruzada é chamado validação cruzada com eliminação de uma observação (leave-one-out cross-validation - LOOCV).

Validação Cruzada

- Na chamada validação cruzada de ordem k (*k-fold cross validation*) o conjunto de dados original é subdividido em dois, sendo um deles utilizado como conjunto de treinamento e o segundo como conjunto de teste. Esse processo é repetido k vezes com conjuntos de treinamento e testes diferentes como esquematizado na Figura 5.
- O correspondente erro de previsão é estimado como

$$VC_{(k)} = \frac{1}{k}\sum_{i=1}^{k} EQM_i. \tag{2}$$

em que

$$EQM_i = \sum (y_j - \widehat{y}_j)^2 / n_i$$

é erro quadrático médio obtido no i-ésimo ajuste, $i = 1, \ldots, k$.
Aqui, y_j, $\widehat{y}_j$ e n_i são, respectivamente, os valores observado e previsto para a j-ésima observação e o número de observações no i-ésimo conjunto de teste. O usual é tomar $k = 5$ ou $k = 10$.

Validação Cruzada

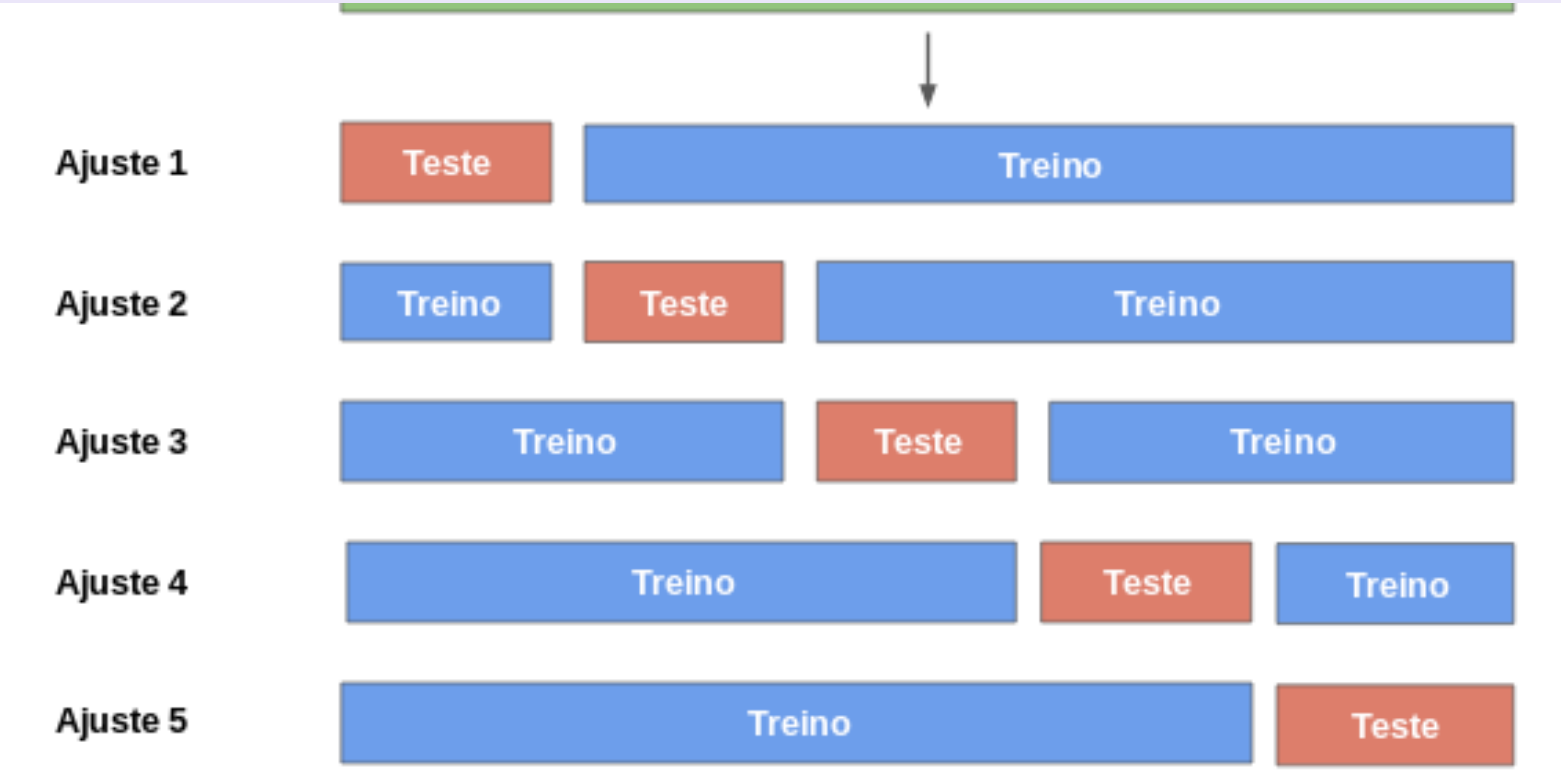


Figura 5: Representação esquemática da divisão dos dados para validação cruzada de ordem k.

Bootstrap

- Com o progresso de métodos computacionais e com capacidade cada vez maior de lidar com conjuntos grandes de dados, o cálculo de erros padrões, vieses etc, pode ser feito sem recorrer a uma teoria, que muitas vezes pode ser muito complicada ou simplesmente não existir.
- Um desses métodos é chamado bootstrap , introduzido por B. Efrom, em 1979. A ideia básica do método bootstrap é re-amostrar o conjunto disponível de dados para estimar um parâmetro θ, com o fim de criar dados replicados. A partir dessas replicações, podemos avaliar a variabilidade de um estimador proposto para θ, sem recorrer a cálculos analíticos.
- Suponha que temos dados $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e queremos estimar a mediana populacional, Md, por meio da mediana amostral $\text{md}(\mathbf{x}) = \text{med}(x_1, \ldots, x_n)$.
- Escolhemos uma amostra aleatória simples, *com reposição*, de tamanho n dos dados. Tal amostra é chamada uma amostra bootstrap e denotada por $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$.

Bootstrap

- Suponha, agora, que geramos B tais amostras independentes, denotadas $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$. Para cada amostra bootstrap, geramos uma réplica bootstrap do estimador proposto, ou seja, de $\text{md}(x)$, obtendo-se

$$\text{md}(\mathbf{x}_1^*), \text{md}(\mathbf{x}_2^*), \ldots, \text{md}(\mathbf{x}_B^*). \tag{3}$$

- Definimos o estimador bootstrap do erro padrão de $\text{md}(\mathbf{x})$ como

$$\widehat{\text{e.p.}}_B(\text{md}) = \left[\frac{\sum_{b=1}^{B} (\text{md}(\mathbf{x}_b^*) - \text{md}(\cdot))^2}{B-1} \right]^{1/2}, \tag{4}$$

com

$$\text{md}(\cdot) = \frac{\sum_{b=1}^{B} \text{md}(\mathbf{x}_b^*)}{B}. \tag{5}$$

Ou seja, o estimador bootstrap do erro padrão da mediana amostral é o desvio padrão amostral do conjunto (3).

Bootstrap

- A questão que se apresenta é: qual deve ser o valor de B, ou seja, quantas amostras bootstrap devemos gerar para estimar erroes padrões de estimadores? A experiência indica que um valor razoável é $B = 200$.
- No caso geral de um estimador $\hat{\theta} = t(\mathbf{x})$, o algoritmo bootstrap para estimar o erro padrão de $\hat{\theta}$ é o seguinte:

[1] Selecione B amostras bootstrap independentes $\mathbf{x}_1^*, \ldots, \mathbf{x}_B^*$, cada uma consistindo de n valores selecionados com reposição de $\mathbf{x}$. Tome $B \approx 200$.

[2] Para cada amostra bootstrap $\mathbf{x}_b^*$ calcule a réplica bootstrap

$$\hat{\theta}^*(b) = t(\mathbf{x}_b^*), \quad b = 1, 2, \ldots, B.$$

[3] O erro padrão de $\hat{\theta}$ é estimado pelo desvio padrão das B réplicas:

$$\widehat{\text{e.p.}}_B = \left[\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}^*(b) - \hat{\theta}^*(\cdot))^2 \right]^{1/2}, \tag{6}$$

com

$$\hat{\theta}^*(\cdot) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^*(b). \tag{7}$$

Classificação

- De modo genérico, vamos designar a variável preditora por X (que pode ser escalar ou vetorial) e a resposta (indicadora de uma classe) por Y.
- Os dados serão indicados por (x_i, y_i), $i = 1, \ldots, n$. A ideia é usar os dados para obter agrupamentos cujos elementos sejam de alguma forma parecidos entre si (com base em alguma medida obtida a partir da variável preditora) e depois utilizar essa medida para classificar um ou mais novos elementos (para os quais dispomos apenas dos valores da variável preditora) em uma das classes.
- Se tivermos d variáveis preditoras e uma resposta dicotômica (*i.e.*, duas classes), um classificador é uma função que mapeia um espaço d-dimensional sobre $\{-1, 1\}$.
- Formalmente, seja (X, Y) um vetor aleatório, de modo que $X \in \mathbb{R}^d$ e $Y \in \{-1, 1\}$. Então, um classificador é uma função $g : \mathbb{R}^d \to \{-1, 1\}$ e a função erro ou risco é a probabilidade de erro, $L(g) = P\{g(X) \neq Y\}$.

Classificação

- A acurácia de um estimador de g, digamos $\widehat{g}$, pode ser medida pelo estimador de $L(g)$, chamado de taxa de erros, que é a proporção de erros gerados pela aplicação de $\widehat{g}$ às observações do conjunto de dados, ou seja,

$$\widehat{L}(\widehat{g}) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \widehat{y}_i), \tag{8}$$

  com $\widehat{y}_i = \widehat{g}(x_i)$ indicando o rótulo (-1 ou 1) da classe prevista por meio de $\widehat{g}$. Se $I(y_i \neq \hat{y}_i) = 0$, a i-ésima observação estará classificada corretamente.
- Sob o enfoque de aprendizado automático (AA), o objetivo é comparar diferentes modelos para identificar aquele com menor taxa de erros.

Classificação

- Nesse contexto, dispomos de um conjunto de dados de treinamento (x_i, y_i), $i = 1, \ldots, n$ e de um conjunto de dados de teste, cujo elemento típico é (x_0, y_0). O interesse é minimizar a taxa de erro de teste associada ao conjunto de observações teste que pode ser estimada por

$$\text{Média}[I(y_0 \neq \widehat{y_0})], \tag{9}$$

  em que a média é calculada relativamente aos elementos do conjunto de dados de teste.
- O classificador (ou modelo) ótimo é aquele que minimiza (9).
- Com o objetivo de classificar os elementos do conjunto de dados, deve-se ajustar o classificador ótimo ao conjunto de dados disponíveis (treinamento e teste)e utilizar a estimativa $\widehat{g}$ daí obtida para classificar os elementos do conjunto de dados para classificação.
- Quando dispomos de apenas um conjunto de dados, podemos recorrer ao processo de validação cruzada para dividi-lo em conjuntos de dados de treinamento e de dados de teste.

Classificação clássica

Neste contexto, podemos ter:

1) Classificação por regressão logística;
2) Classificação bayesiana;
3) Função discriminante linear de Fisher;
4) Classificador K-vizinho mais próximo.
5) Outras propostas

Algumas medidas

Considere a tabela (matriz de confusão):

		Condição Verdadeira	
		Positiva	Negativa
Condição Prevista	Positiva	n_{11} (TP)	n_{12} (FP)
	Negativa	n_{21} (FN)	n_{22} (TN)
		CP	CN

TP=True positive, FP= False positive, FN= False negative, TN= True negative

TPR= $\frac{TP}{TP+FN} = \frac{n_{11}}{n_{11}+n_{21}}$ (sensibilidade) (True positive rate)

TNR= $\frac{TN}{FP+TN} = \frac{n_{22}}{n_{12}+n_{22}}$ (especificidade) (True negative rate)

Prevalence= $\frac{CP}{CP+CN}$

PPV= $\frac{TP}{TP+FP}$ (Positive predictive value, precision)

NPV= $\frac{TN}{FN+TN}$ (Negative predictive value)

FDR= $\frac{FP}{TP+FP}$ (False discovery rate)

Accuracy=ACC= $\frac{TP+TN}{CP+CN}$, Balanced accuracy= BA= $\frac{TPR+TNR}{2}$

Classificação por regressão logística - RL

- **Exemplo**. Os dados da Tabela 1 são extraídos de um estudo realizado no Hospital Universitário da Universidade de São Paulo com o objetivo de avaliar se algumas medidas obtidas ultrassonograficamente poderiam ser utilizadas como substitutas de medidas obtidas por métodos de ressonância magnética, considerada como padrão áureo, para avaliação do deslocamento do disco da articulação temporomandibular (doravante referido simplesmente como disco).
- Distâncias cápsula-côndilo (em mm) com boca aberta ou fechada (referidas, respectivamente, como distância aberta ou fechada no restante do texto) foram obtidas ultrassonograficamente de 104 articulações e o disco correspondente foi classificado como deslocado (1) ou não (0) segundo a avaliação por ressonância magnética. A variável resposta é o *status* do disco (1 = deslocado ou 0 = não).
- Consideremos um modelo logístico para a chance de deslocamento do disco, tendo apenas a distância aberta como variável explicativa: Nesse contexto, o modelo

$$\log[\theta(x_i;\alpha,\beta)]/[1-\theta(x_i;\alpha,\beta)] = \alpha + x_i\beta \qquad (10)$$

$i = 1, \ldots, 104$.

RL-Exemplo

Tabela: Parte dos Dados de um estudo odontológico

Dist aberta	Dist fechada	Desloc disco	Dist aberta	Dist fechada	Desloc disco	Dist aberta	Dist fechada	Desloc disco
2.2	1.4	0	0.9	0.8	0	1.0	0.6	0
2.4	1.2	0	1.1	0.9	0	1.6	1.3	0
2.6	2.0	0	1.4	1.1	0	4.3	2.3	1
3.5	1.8	1	1.6	0.8	0	2.1	1.0	0
1.3	1.0	0	2.1	1.3	0	1.6	0.9	0
2.8	1.1	1	1.8	0.9	0	2.3	1.2	0
1.5	1.2	0	2.4	0.9	0	2.4	1.3	0
2.6	1.1	0	2.0	2.3	0	2.0	1.1	0
1.2	0.6	0	2.0	2.3	0	1.8	1.2	0
1.7	1.5	0	2.4	2.9	0	1.4	1.9	0
1.3	1.2	0	2.7	2.4	1	1.5	1.3	0
1.2	1.0	0	1.9	2.7	1	2.2	1.2	0
4.0	2.5	1	2.4	1.3	1	1.6	2.0	0
1.2	1.0	0	2.1	0.8	1	1.5	1.1	0
3.1	1.7	1	0.8	1.3	0	1.2	0.7	0
2.6	0.6	1	0.8	2.0	1	1.5	0.8	0
1.8	0.8	0	0.5	0.6	0	1.8	1.1	0
1.2	1.0	0	1.5	0.7	0	2.3	1.6	1
1.9	1.0	0	2.9	1.6	1	1.2	0.4	0
1.2	0.9	0	1.4	1.2	0	1.0	1.1	0
1.7	0.9	1	3.2	0.5	1	2.9	2.4	1
1.2	0.8	0	1.2	1.2	0	2.5	3.3	1

Dist aberta: distância cápsula-côndilo com boca aberta (mm)
Dist fechada: distância cápsula-côndilo com boca fechada (mm)
Desloc disco: deslocamento do disco da articulação temporomandibular (1=sim, 0=não)

RL–Exemplo

- No modelo, $\theta(x_i; \alpha, \beta)$ representa a probabilidade de deslocamento do disco quando o valor da distância aberta é x_i, α denota o logaritmo da chance de deslocamento do disco quando a distância aberta tem valor $x_i = 0$ e β é interpretado como a variação no logaritmo da chance de deslocamento do disco por unidade de variação da distância aberta.
- Consequentemente, a razão de chances do deslocamento do disco correspondente a uma diferença de d unidades da distância aberta será $\exp(d \times \beta)$. Como não temos dados correspondentes a distâncias abertas menores que 0,5, convém substituir os valores x_i por valores "centrados", ou seja por $x_i^* = x_i - x_0$.
- Uma possível escolha para x_0 é o mínimo de x_i, que é 0,5. Essa transformação na variável explicativa altera somente a interpretação do parâmetro α que passa a ser o logaritmo da chance de deslocamento do disco quando a distância aberta tem valor $x_i = 0,5$.

RL–Exemplo–Uso do R

```
Call:
glm(formula = deslocamento ~ (distanciaAmin), family = binomial,
    data = disco)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5240  -0.4893  -0.3100   0.1085   3.1360

Coefficients:
             Estimate  Std. Error z value Pr(>|z|)
(Intercept)   -5.8593     1.1003  -5.325  1.01e-07 ***
distanciaAmin  3.1643     0.6556   4.827  1.39e-06 ***
---


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.11  on 103  degrees of freedom
Residual deviance:  71.60  on 102  degrees of freedom
AIC: 75.6

Number of Fisher Scoring iterations: 6
```

RL–Exemplo

- Estimativas (com erros padrões entre parênteses) dos parâmetros desse modelo ajustado por máxima verossimilhança aos dados da Tabela 1, são, $\widehat{\alpha} = -5,86\ (1,10)$ e $\widehat{\beta} = 3,16\ (0,66)$ e então, segundo o modelo, uma estimativa da chance de deslocamento do disco para articulações com distância aberta $x = 0,5$ (que corresponde à distância aberta transformada $x^* = 0,0$) é $\exp(-5,86) = 0,003$.
- Um intervalo de confiança (95%) para essa chance pode ser obtido exponenciando os limites (LI e LS) do intervalo para o parâmetro α, nomeadamente,

$$\begin{aligned} LI &= \exp[\widehat{\alpha} - 1,96EP(\widehat{\alpha})] = \exp(-5,86 - 1,96 \times 1,10) = 0,000 \\ LS &= \exp[\widehat{\alpha} + 1,96EP(\widehat{\alpha})] = \exp(-5,86 + 1,96 \times 1,10) = 0,025. \end{aligned}$$

- Os limites de um intervalo de confiança para a razão de chances correspondentes a um variação de uma unidade no valor da distância aberta podem ser obtidos de maneira similar e são 6,55 e 85,56.
- Substituindo os parâmetros α e β por suas estimativas $\widehat{\alpha}$ e $\widehat{\beta}$ em (10) podemos estimar a probabilidade de sucesso (deslocamento do disco, no exemplo sob investigação).

RL–Exemplo

- Por exemplo, para uma articulação cuja distância aberta seja 2,1 (correspondente à distância aberta transformada igual a 1,6), a estimativa dessa probabilidade é

$$\widehat{\theta} = \exp(-5,86 + 3,16 \times 1,6)/[1 + \exp(-5,86 + 3,16 \times 1,6)] = 0,31.$$

- Lembrando que o objetivo do estudo é substituir o processo de identificação de deslocamento do disco realizado via ressonância magnética por aquele baseado na medida da distância aberta por meio de ultrassonografia, podemos estimar as probabilidades de sucesso para todas as articulações e identificar um ponto de corte d_0 segundo o qual, distâncias abertas com valores acima dele sugerem decidirmos pelo deslocamento do disco e distâncias abertas com valores abaixo dele sugerem a decisão oposta.
- Obviamente, não esperamos que todas as decisões tomadas dessa forma sejam corretas e consequentemente, a escolha do ponto de corte deve ser feita com o objetivo de minimizar os erros (decidir pelo deslocamento quando ele não existe ou *vice versa*).

RL–Exemplo

Nesse contexto, um contraste entre as decisões tomadas com base em um determinado ponto de corte d_0 e o padrão áureo definido pela ressonância magnética para todas as 104 articulações pode ser resumido por meio da Tabela 2, em que as frequências da diagonal principal correspondem a decisões corretas e aquelas da diagonal secundária às decisões erradas.

Tabela: Frequência de decisões para um ponto de corte d_0

		Deslocamento real do disco	
		sim	não
Decisão baseada na distância aberta d_0	sim	n_{11}	n_{12}
	não	n_{21}	n_{22}

RL–Exemplo

- O quociente $n_{11}/(n_{11} + n_{21})$ é conhecido como sensibilidade do processo de decisão e é uma estimativa da probabilidade de decisões corretas quando o disco está realmente deslocado.
- O quociente $n_{22}/(n_{12} + n_{22})$ é conhecido como especificidade do processo de decisão e é uma estimativa da probabilidade de decisões corretas quando o disco realmente não está deslocado. A situação ideal é aquela em que tanto a sensibilidade quanto a especificidade do processo de decisão são iguais a 100%.
- O problema a resolver é determinar o ponto de corte d_{max} que gere o melhor equilíbrio entre sensibilidade e especificidade. Com essa finalidade, podemos construir tabelas com o mesmo formato da Tabela 2 para diferentes pontos de corte e um gráfico cartesiano entre a sensibilidade e especificidade obtida de cada uma delas.
- Esse gráfico, conhecido como curva ROC (do termo inglês Receiver Operating Characteristic) gerado para os dados da Tabela 1 está apresentado na Figura 1.

RL–Exemplo

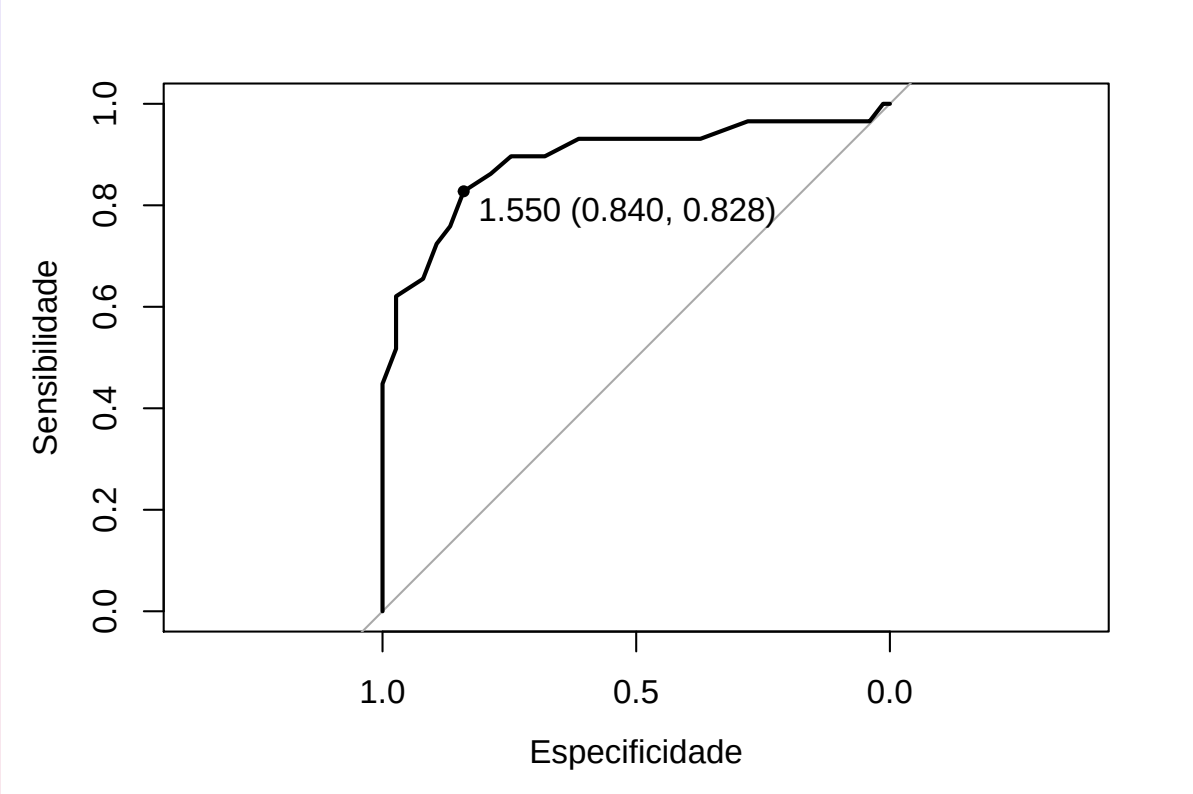


Figura: Curva ROC para os dados da Tabela 1 baseada no modelo (10) com distância aberta como variável explicativa.

RL–Exemplo

- O ponto de corte ótimo é aquele mais próximo do vértice superior esquerdo (em que tanto a sensibilidade quanto a especificidade seriam iguais a 100%).
- Para o exemplo, esse ponto está salientado na Figura 1 e corresponde à distância aberta com valor $d_{max} = 2,05$ $(= 1,55 + 0,5)$. A sensibilidade e a especificidade associadas à decisão baseada nesse ponto de corte, são, respectivamente, 83% e 84% e as frequências de decisões corretas estão indicadas na Tabela 3.

Tabela: Frequência de decisões para um ponto de corte para distância aberta $d_{max} = 2,05$

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	24	12
distância aberta $d_{max} = 2,05$	não	5	63

RL–Exemplo

- Com esse procedimento de decisão a porcentagem de acertos (**acurácia**) é 84% $[= (24 + 63)/104]$. A porcentagem de **falsos positivos** é 17% $[= 5/(5 + 29)]$ e a porcentagem de **falsos negativos** é 16% $[= 12/(12 + 63)]$.
- Um gráfico de dispersão com o correspondente ponto de corte baseado apenas na distância aberta está apresentado na Figura 2 com símbolos vermelhos indicando casos com deslocamento do disco e em preto indicando casos sem deslocamento. Os valores de ambas as distâncias foram ligeiramente alterados para diminuir a superposição nos pontos.

RL–Exemplo

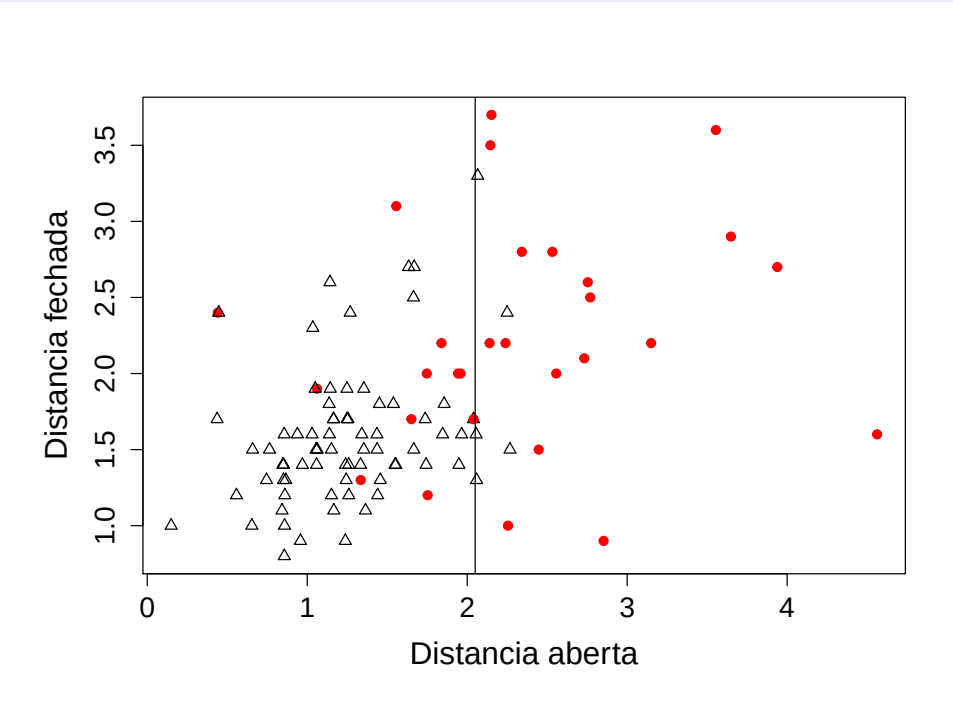


Figura: Gráfico de dispersão para os dados da Tabela 1 com ponto de corte baseado apenas na distância aberta.

RL–Exemplo

Uma análise similar, baseada na distância fechada (transformada por meio da subtração de deu valor mínimo, 0,4) gera a curva ROC apresentada na Figura 3 e frequências de decisões apresentada na Tabela 4.

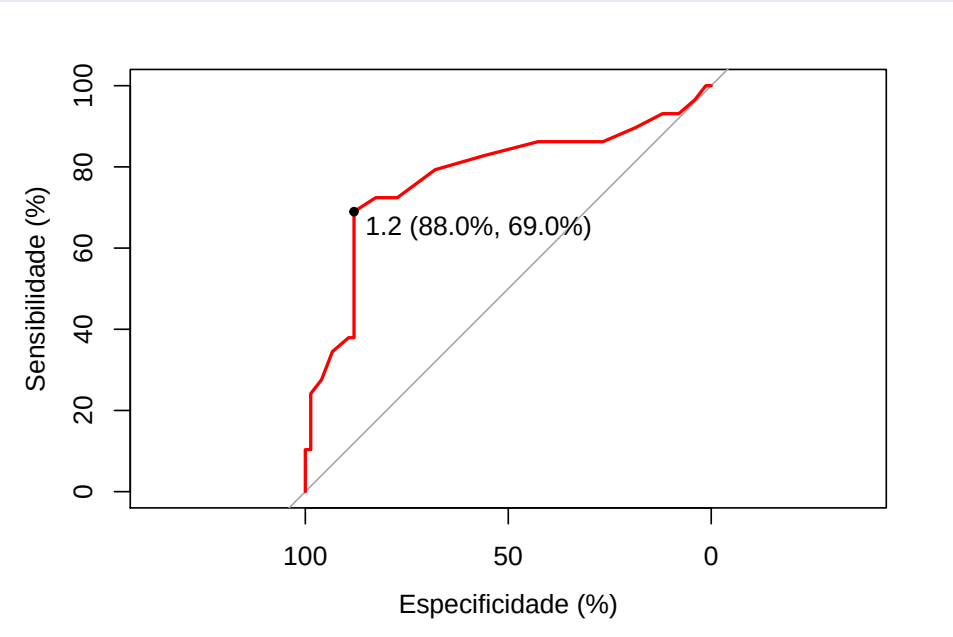


Figura: Curva ROC para os dados da Tabela 1 baseada no modelo (10) com distância fechada como variável explicativa.

RL–Exemplo

Tabela: Frequência de decisões para um ponto de corte para distância fechada $d_{max} = 1,6$

		Deslocamento real do disco	
		sim	não
Decisão baseada na	sim	20	9
distância fechada $d_{max} = 1,6$	não	9	66

RL–Exemplo

- A acurácia associada a processo de decisão baseado apenas na distância fechada, 83% $[= (20 + 66)/104]$ é praticamente igual àquela obtida com base apenas na distância aberta; no entanto aquele processo apresenta um melhor equilíbrio entre sensibilidade e especificidade (83% e 84%, respectivamente, *versus* 69% e 88%).
- Se quisermos avaliar o processo de decisão com base nas observações das distâncias aberta e fechada simultaneamente, podemos considerar o modelo

$$\log[\theta(x_i; \alpha, \beta, \gamma)]/[1 - \theta(x_i; \alpha, \beta, \gamma)] = \alpha + x_i\beta + w_i\gamma \qquad (11)$$

$i = 1, \ldots, 104$ em que w_i corresponde à distância fechada observada na i-ésima articulação.

RL–Exemplo

- Neste caso, γ corresponde à razão entre a chance de deslocamento do disco para articulações com distância fechada $w+1$ e a chance de deslocamento do disco para articulações com distância fechada w para aquelas com mesmo valor da distância aberta; uma interpretação similar vale para o parâmetro β.
- Estimativas dos parâmetros (com erros padrões entre parênteses) do modelo (11) obtidas após a transformação das variáveis explicativas segundo o mesmo figurino adotado nas análises univariadas são $\widehat{\alpha} = -6,38\ (1,19)$, $\widehat{\beta} = 2,83\ (0,67)$ e $\widehat{\gamma} = 0,98\ (0,54)$.
- A estimativa do parâmetro γ é apenas marginalmente significativa, ou seja a inclusão da variável explicativa distância fechada não acrescenta muito poder de discriminação além daquele correspondente à distância aberta.

RL–Exemplo

- Uma das razões para isso é que as duas variáveis são correlacionadas (com coeficiente de correlação de Pearson igual a 0,46. A determinação de pontos de corte para modelos com duas ou mais variáveis explicativas é bem mais complexa do que no caso univariado e não será abordada neste texto.
- Para efeito de comparação com as análises anteriores, as frequências de decisões obtidas com os pontos de corte utilizados naquelas estão dispostas na Tabela 5, e correspondem a uma sensibilidade de 62%, especificidade de 97% e acurácia de 88%.

Tabela: Frequência de decisões correspondentes a pontos de corte $d_{max} = 2,05$ para distância aberta e $d_{max} = 1,6$ para distância fechada

		Deslocamento real do disco	
		sim	não
Decisão baseada em	sim	18	2
ambas as distâncias	não	11	73

RL–Exemplo

- Numa segunda análise, agora sob o paradigma de aprendizado automático (AA), a escolha do modelo ótimo é baseada apenas nas porcentagens de classificação correta (acurácia) obtidas por cada modelo num conjunto de dados de teste a partir de seu ajuste a um conjunto de dados de treinamento. Como neste caso não dispomos desses conjuntos *a priori*, podemos recorrer à técnica de **validação cruzada**.
- Neste exemplo, utilizamos validação cruzada de ordem 5 com 5 repetiçoes (VC5/5), em que o conjunto de dados é dividido em dois, cinco vezes, gerando cinco conjuntos de dados de treinamento e de teste. A análise é repetida cinco vezes em cada conjunto e a a acurácia média obtida das 25 análises serve de base para a escolha do melhor modelo.
- Comparamos quatro modelos de regressão logística, os dois primeiros com apenas uma das variáveis preditoras (distância aberta ou distância fechada), o terceiro com ambas incluídas aditivamente e o último com ambas as distâncias e sua interação. A análise pode ser concretizada por meio do pacote caret.

RL–Exemplo

Os resultados estão dispostos na Tabela 6 tanto para validação cruzada VC5/5 quanto para validação cruzada LOOCV.

Tabela: Acurácia obtida por validação cruzada para as regressões logísticas ajustados as dados do Exemplo 8.1

Modelo	Variáveis	Acurácia VC5/5	Acurácia LOOCV
1	Distância aberta	84,8 %	84,6 %
2	Distância fechada	75,2 %	74,0 %
3	Ambas (aditivamente)	85,7 %	85,6 %
4	Ambas + Interação	83,6 %	83,6 %

RL–Exemplo

- Com ambos os critérios, o melhor modelo é aquele que inclui as duas variáveis preditoras de forma aditiva. Para efeito de classificar uma nova observação (para a qual só dispomos dos valores das variáveis preditoras, o modelo selecionado deve ser ajustado ao conjunto de dados original (treinamento + teste) para obtenção dos coeficientes do classificador.
- Os comandos e a saída associada ao ajuste desse modelo aos 5 conjuntos de dados gerados para validação cruzada e no conjunto completo seguem.
- A seleção obtida por meio do AA corresponde ao modelo (11). Embora a variável Distância fechada seja apenas marginalmente significativa, sua inclusão aumenta a proporção de acertos (acurácia) de 84% no modelo que inclui apenas Distância aberta para 86%.
- A estatísca Kappa apresentada juntamente com a acurácia serve para avaliar a concordância entre o processo de classificação e a classificação observada (veja a Seção 4.2).

RL–Exemplo

```
> set.seed(369321)
> train_control =
                trainControl(method="repeatedcv", number=5, repeats=5)
> model3 = train(deslocamento ~ distanciaAmin + distanciaFmin,
           data=disco, method="glm", family=binomial,
           trControl=train_control)
> model3
Generalized Linear Model

104 samples
  2 predictor
  2 classes: 0, 1

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 5 times)
Summary of sample sizes: 83, 83, 84, 83, 83, 83, ...
Resampling results:

  Accuracy   Kappa
  0.8573333  0.6124102

> disco$predito3 = predict(model3, newdata=disco, type="raw")
> summary(model3$finalModel)
```

RL–Exemplo

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.82771  -0.45995  -0.28189   0.07403   2.82043

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.3844     1.1932  -5.351 8.76e-08 ***
distanciaAmin  2.8337     0.6676   4.245 2.19e-05 ***
distanciaFmin  0.9849     0.5383   1.830   0.0673 .


(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 123.107  on 103  degrees of freedom
Residual deviance:  67.991  on 101  degrees of freedom
AIC: 73.991
Number of Fisher Scoring iterations: 6

> table(disco$deslocamento, disco$predito3)
     0  1
  0 72  3
  1 12 17
```

Como a divisão do conjunto original nos subconjuntos de treinamento e de teste envolve uma escolha aleatória, os resultados podem diferir (em geral de forma desprezável) para diferentes aplicações dos mesmos comandos, a não ser que se especifique a semente do processo aleatório de divisão por meio do comando set.seed().

Kappa de Cohen

(i) Estatística que mede a concordância entre dois avaliadores, cada um classificando N itens em C categorias mutuamente exclusivas.

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

na qual p_0 é o valor de concordância observado relativo entre os avaliadores e p_e é a probabilidade sob a hipótese de concordância usando os dados observados para calcular a probabilidade de que cada avaliador veja cada categoria aleatoriamente.

(ii) Se $\kappa = 1$ os avaliadores concordam totalmente e $\kappa = 0$ se discordam completamente.

(iii) No caso de classificação, temos

$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$$

Kappa de Cohen

Exemplo: 50 pessoas são avaliadas para receber uma *grant* por dois avaliadores.

A \ B	Sim	Não	Total
Sim	20	5	25
Não	10	15	25
Total	30	20	50

$$p_0 = \frac{20+15}{50} = 0,7.$$

$$p_{\text{Sim}} = \frac{20+5}{50} \cdot \frac{20+10}{50} = 0.5 \times 0.6 = 0.3$$

Esta é a probabilidade esperada que ambos digam Sim

$$p_{\text{Não}} = \frac{10+15}{50} \cdot \frac{5+15}{50} = 0.5 \times 0.4 = 0.2.$$

Esta é a probabilidade esperada que ambos digam Não

$$p_e = p_{\text{Sim}} + p_{\text{Não}} = 0.5$$

$$\kappa = \frac{0.7 - 0.5}{1 - 0.5} = 0.4$$

Kappa de Cohen

No exemplo de RL: TP=32, TN=17, FN=12, FP=3, o software forneceu $\kappa = 0,612$.

Usando a fórmula (iii) obtemos

$$\kappa = \frac{2 \times (72 \times 17 - 12 \times 3)}{75 \times 20 + 84 \times 29} = \frac{2376}{3936} = 0,604.$$

Área sob a curva ROC

- Seja $\mathcal{F}(\mathbf{X}(\omega))$ um *classificador*, $\mathbf{X}(\omega)$ é um vetor de v.a.'s de dimensão p. Dividimos o universo Ω em duas classes:

$$\mathcal{N}_\lambda = \{\omega \in \Omega : \mathcal{F}(\mathbf{X}(\omega)) \leq \lambda\} \quad \text{e } \mathcal{P}_\lambda = \mathcal{N}_\lambda^c. \tag{12}$$

- O propósito de um teste é discriminar os membros de Ω entre as duas classes nas quais temos interesse.Vamos chamá-las $\mathcal{H}$ e $\mathcal{U} = \mathcal{H}^c$.
- O teste ideal seria aquele para o qual existe um λ tal que $\mathcal{N}_\lambda = \mathcal{H}$.
- Suponha que fdp da v.a. $\mathcal{F}$ seja dada pela função contínua $f(\mathbf{x})$ para $\mathbf{x} \in \mathbb{R}^p$. Sejam $f_{\mathcal{H}}(\mathbf{x})$ e $f_{\mathcal{U}}(\mathbf{x})$ as fdp's de $\mathcal{F}$ condicionais às classes $\mathcal{H}$ e $\mathcal{U}$, respectivamente, supostas contínuas.
- Agora se p_H é a probabilidade do evento $\mathcal{H}$ (equivalentemente, $p_H(\mathbf{x})$é a probabilidade de que um objeto com vetor $X(\omega) = \mathbf{x}$ pertença à classe $\mathcal{H}$), temos

$$f(x) = p_H(\mathbf{x}) f_{\mathcal{H}}(x) + (1 - p_H) f_{\mathcal{U}}(x). \tag{13}$$

Área sob a curva ROC

- A probabilidade de associar um particular ω de $\mathcal{P}_\lambda$ à classe $\mathcal{U}$ é

$$\sigma(\lambda) = \int_\lambda^\infty f_{\mathcal{U}}(x)\, dx. \tag{14}$$

  Chame esse número de sensibilidade da partição $(\mathcal{N}_\lambda, \mathcal{P}_\lambda)$, a sensibilidade associada à λ ou a sensibilidade do teste.

- E a probabilidade de associar um certo ω de $\mathcal{N}_\lambda$ à classe $\mathcal{H}$ é chamada a especificidade da partição $(\mathcal{N}_\lambda, \mathcal{P}_\lambda)$, a especificidade associada a λ ou a especificidade do teste, dada por

$$\varepsilon(\lambda) = \int_{-\infty}^\lambda f_{\mathcal{H}}(x)\, dx. \tag{15}$$

Área sob a curva ROC

- O gráfico da relação

$$\begin{aligned}\mathcal{R} &= \{(1-\int_{-\infty}^{\lambda} f_{\mathcal{H}}(x)\,dx, \int_{\lambda}^{\infty} f_{\mathcal{U}}(x)\,dx) : \lambda \in \mathbb{R}\} \\ &= \{(\int_{\lambda}^{\infty} f_{\mathcal{H}}(x)\,dx, \int_{\lambda}^{\infty} f_{\mathcal{U}}(x)\,dx) : \lambda \in \mathbb{R}\} \qquad (16)\end{aligned}$$

  é a curva ROC (Receiver Operating Characteristic) do teste $\mathcal{F}$.
- É comum definir a acurácia de um teste como a área sob a curva ROC.
- Quando as densidade $f_{\mathcal{H}}(x)$ and $f_{\mathcal{U}}(x)$ são gaussianas, dizemosque a curva ROC é normal.

Área sob a curva ROC

- Pode-se mostrar que a área sob a curva ROC é dada por

$$\mathcal{A}_\gamma = 1 - \int_{-\infty}^{\infty} \left(\int_{\lambda}^{\infty} f_{\mathcal{H}}(t)\, dt \right) \phi_{\mathcal{U}}(\lambda)\, d\lambda. \tag{17}$$

- Após alguma s manipulações (teorema de Fubini) obtemos

$$\mathcal{A}_\gamma = \mathbb{E}_{\mathcal{H}}[\sigma] = \mathbb{E}_{\mathcal{U}}[\varepsilon]. \tag{18}$$

- Aqui $\mathbb{E}_{\mathcal{H}}$ e $\mathbb{E}_{\mathcal{U}}$ são os valores esperados com respeito às densidades $f_{\mathcal{H}}$ e $f_{\mathcal{U}}$, respectivamente.
- Em outras palavras, a acurácia é a média condicional da sensibilidade sob $\mathcal{H}$ ou a média condiconal da especificidade sob $\mathcal{U}$.

Área sob a curva ROC

- Se chamarmos as f.d.'s de f, $f_{\mathcal{H}}$ e $f_{\mathcal{U}}$ de F, $F_{\mathcal{H}}$ e $F_{\mathcal{U}}$, respectivamente, pode-se mostrar que

$$\mathcal{A}_\gamma = \frac{2\mathbb{E}[\sigma] - (1 - p_H)}{2p_H} = \frac{2\mathbb{E}[\varepsilon] - p_H}{2(1 - p_H)}, \tag{19}$$

  em que $\mathbb{E}$ é a média incondicional com respeito a f.

- Suponha que as fdp's sejam

$$f_{\mathcal{H}}(x) = \frac{1}{\sqrt{2\pi}\sigma_H} \exp\left[-\frac{1}{2}\left(\frac{x - \mu_H}{\sigma_H}\right)^2\right], \quad f_{\mathcal{U}}(x) = \frac{1}{\sqrt{2\pi}\sigma_U} \exp\left[-\frac{1}{2}\left(\frac{x - \mu_U}{\sigma_U}\right)^2\right]. \tag{20}$$

Área sob a curva ROC

- **Proposição** Dados dois testes A e B para os quais as distribuições da sensibilidade e especificidade sejam normais com parâmetros μ_U^A, σ_U^A e μ_H^A, σ_H^A, respectivamente, para o teste A, e μ_U^B, σ_U^B and μ_H^B, σ_H^B, respectivamente, para o teste B, então o teste com a maior área sob a curva ROC é aquele para o qual $\frac{\mu_U - \mu_H}{\sigma_H}$ (ou, alternativamente, $\frac{\mu_H - \mu_U}{\sigma_U}$) é o maior.
- **Example.** Suponha que tenhamos um teste A with $\sigma_U^A = \sigma_H^A = 1$ e $\mu_U^A - \mu_H^A = 1$ e outro teste B with $\sigma_U^B = \sigma_H^B = 0.999$ e $\mu_U^B - \mu_H^B = 1.001$. Então o teste B é melhor que o teste A.

Área sob a curva ROC

- De (18) obtemos um estimador da área, substituindo as distribuições tóricas pelas amostrais:

$$\hat{A}_\gamma = \int_{-\infty}^{\infty} \hat{F}_{\mathcal{H}}(\lambda)\hat{f}_{\mathcal{U}}(\lambda)d\lambda. \tag{21}$$

- Obtemos

$$\hat{A}_\gamma = \frac{S_0 - n_0(n_0+1)/2}{n_0 n_1}, \tag{22}$$

onde S_0 é a soma dos postos dos valores da classe teste $\mathcal{H}$, n_0 e n_1 são os números de valores das classes $\mathcal{H}$ e $\mathcal{U}$, respectivamente.

- Pode ser mostrado que este estimador da AUC é equivalente ao teste de Mann-Whitney-Wilcoxon e é relacionado com o coeficiente de Gini, G, a saber $2AUC = G + 1$.

Referências

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**, 1–22.

Hastie, T., Tibshirani, R. and Wainwright, M. (2015). *Statistical Learning with Sparcity*. Chapman and Hall.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2017). *Introduction to Statistical Learning*. Springer.

Morettin, P. A. e Singer, J. M. (2024). *Estatística e Ciência de Dados*. 2a. Edição. LTC.

Prandini, J.C., Morettin, P.A. and Chiann, C. (2024). The area under the normal ROC curve. *São Paulo Journal of Mathematical Sciences*, **18**, 1628–1649.