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 8

24 de abril de 2023

Sumário

ADL

Podemos ter:

1) Classificador de Bayes
2) Classificador linear de Fisher
3) Classificador do vizinho mais próximo
4) Outras propostas

ADL: classificador de Bayes

- Consideremos um conjunto de dados, $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, em que $\mathbf{x}_i$ representa os valores de p variáveis preditoras (explicativas) e y_i representa o valor de uma variável resposta indicadora da classe a que o i-ésimo elemento desse conjunto pertence.
- Seja π_k a probabilidade *a priori* de que um elemento com valor das variáveis preditoras $\mathbf{x} = (x_1, \ldots, x_p)$ pertença à classe C_k, $k = 1, \ldots, K$ e seja $f_k(\mathbf{x})$ a função densidade de probabilidade da variável preditora $\mathbf{X}$ para valores $\mathbf{x}$ associados a elementos dessa classe. Por um abuso de notação escrevemos $f_k(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}|Y = k)$ (que a rigor só vale no caso discreto).
- Pelo teorema de Bayes,

$$P(Y = k|\mathbf{X} = \mathbf{x}) = p_k(\mathbf{x}) = \frac{\pi_k f_k(\mathbf{x})}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(\mathbf{x})}, \quad k = 1, \ldots, K, \qquad (1)$$

é a probabilidade a posteriori de que um elemento com valor das variáveis preditoras igual a $\mathbf{x}$ pertença à k-ésima classe. Para calcular essa probabilidade é necessário conhecer π_k e $f_k(\mathbf{x})$; em muitos casos, supõe-se que para os elementos da k-ésima classe, os valores de $\mathbf{X}$ tenham uma distribuição $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, ou seja, com média que depende da classe k e matriz de covariâncias comum a todas as classes.

ADL: classificador de Bayes

- Suponha que $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$. Uma **regra de classificação**, R, consiste em dividir $\mathcal{X}$ em K regiões disjuntas $\mathcal{X}_1, \ldots, \mathcal{X}_K$, tal que se $\mathbf{x} \in \mathcal{X}_k$, o elemento correspondente é classificado em C_k.
- A probabilidade (condicional) de classificação incorreta, *i.e.*, de classificar um elemento com valor das variáveis preditoras $\mathbf{x}$ em C_k, quando de fato ele pertence a C_j, $j \neq k$ usando a regra R, é

$$p(C_k|C_j, R) = \int_{\mathcal{X}_k} f_j(\mathbf{x}) d\mathbf{x}. \tag{2}$$

Se $k = j$ em (2), obtemos a probabilidade de classificação correta do elemento com valor das variáveis preditoras $\mathbf{x}$ em C_k.

ADL: classificador de Bayes

- Em muitos casos é possível incluir um **custo** de classificação incorreta, denotado por $Q(C_k|C_j)$ no procedimento de classificação. Usualmente, esses custos não são iguais e admite-se que $Q(C_k|C_k) = 0$, $k = 1, \ldots, K$.
- O custo médio de classificação incorreta segundo a regra R é dado por

$$\delta(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \left[\sum_{k=1, j \neq k}^{K} p(C_j|C_k, R) Q(C_j|C_k) \right]. \tag{3}$$

- O Classificador de Bayes é obtido por meio da minimização desse custo médio, supondo os custos de classificação incorreta iguais, ou seja, o elemento com valor das variáveis preditoras $\mathbf{x}$ deve ser classificado em C_k, se

$$\delta_k(\mathbf{x}) = \sum_{k=1, j \neq k}^{K} \pi_j f_j(\mathbf{x}) \tag{4}$$

for mínima, $k = 1, \ldots K$.

ADL: classificador de Bayes

- Minimizar (4) é equivalente a classificar $\mathbf{x}$ em C_k se

$$\pi_k f_k(\mathbf{x}) = max_{1 \leq j \leq K}[\pi_j f_j(\mathbf{x})], \tag{5}$$

pois devemos excluir a k-ésima parcela de (4) que seja máxima, relativamente a todas as possíveis exclusões de parcelas.

- Em particular, se $K = 2$, elementos com valor das variáveis preditoras igual a $\mathbf{x}$ devem ser classificados em C_1 se

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{\pi_2}{\pi_1}, \tag{6}$$

e em C_2, caso contrário. Veja Johnson e Wichern (1998) e Ferreira (2011), para detalhes.

ADL: classificador de Bayes

- Suponha o caso $K = 2$ com variáveis $\mathbf{x}$ seguindo distribuições normais, com médias $\boldsymbol{\mu}_1$ para elementos da classe C_1, $\boldsymbol{\mu}_2$ para elementos da classe C_2 e matriz de covariâncias $\boldsymbol{\Sigma}$ comum.
  Usando (1), obtemos

$$P(Y = k|\mathbf{X} = xb) = \frac{\pi_k \exp\{-(\mathbf{x} - \mu_k)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu_k)/2\}}{\sum_{\ell=1}^{K} \pi_\ell \exp\{-(\mathbf{x} - \mu_\ell)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu_\ell)/2\}}, \quad k = 1, 2. \tag{7}$$

- Então, elementos com valores das variáveis preditoras iguais a $\mathbf{x}$ são classificados em C_1 se

$$\mathbf{d}^\top \mathbf{x} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} \geq \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \log(\pi_2/\pi_1). \tag{8}$$

em que $\mathbf{d} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ contém os coeficientes da função discriminante.

ADL: classificador de Bayes

- No caso geral ($K \geq 2$), o classificador de Bayes associa um elemento com valor das variáveis preditoras igual a $\mathbf{x}$ à classe para a qual

$$\delta_k(\mathbf{x}) = \boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k \tag{9}$$

  for **máxima**.

- Em particular, para $p = 1$, devemos maximizar

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k. \tag{10}$$

- Quando há apenas duas classes, C_1 e C_2, um elemento com valor da variável preditora igual a x deve ser classificado na classe C_1 se

$$dx = \frac{\mu_1 - \mu_2}{\sigma^2}x \geq \frac{\mu_1^2 - \mu_2^2}{2\sigma^2} + \log\frac{\pi_2}{\pi_1} \tag{11}$$

  e na classe C_2 em caso contrário.

ADL: classificador de Bayes

- As fronteiras de Bayes [valores de $\mathbf{x}$ para os quais $\delta_k(\mathbf{x}) = \delta_\ell(\mathbf{x})$] são obtidas como soluções de

$$\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log \pi_k = \boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_\ell^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_\ell + \log \pi_\ell, \quad (12)$$

  para $k \neq \ell$.

- No paradigma bayesiano, os termos utilizados para cálculo das probabilidades *a posteriori* (1) são conhecidos, o que na prática não é realista. No caso $p = 1$ pode-se aproximar o classificador de Bayes substituindo π_k, μ_k, $k = 1\ldots, K$ e σ^2 pelas estimativas

$$\widehat{\pi}_k = n_k/n$$

  em que n_k corresponde ao número dos n elementos do conjunto de dados de treinamento pertencentes à classe k,

$$\overline{x}_k = \frac{1}{n_k}\sum_{i:y_i=k} x_i, \quad \text{e} \quad S^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \widehat{\mu}_k)^2.$$

ADL: classificador de Bayes

- Com esses estimadores, a fronteira de decisão de Bayes corresponde a solução de

$$(\overline{x}_1 - \overline{x}_2)x = (\overline{x}_1^2 - \overline{x}_2^2)/2 + [\log(\widehat{\pi}_2/\widehat{\pi}_1)]S^2. \quad (13)$$

- No caso $K = 2$ e $p \geq 2$, os parâmetros $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ são desconhecidas e têm que ser estimadas a partir de amostras das variáveis preditoras associadas aos elementos de C_1 e C_2. Com os dados dessas amostras, podemos obter estimativas $\overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \mathbf{S}_1$ e $\mathbf{S}_2$, das respectivas médias e matrizes de covariâncias.
- Uma estimativa não enviesada da matriz de covariâncias comum $\boldsymbol{\Sigma}$ é

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}. \quad (14)$$

- Quando $\pi_1 = \pi_2$, elementos com valor das variáveis preditoras igual a $\mathbf{x}$ são classificados em C_1 se

$$\widehat{\mathbf{d}}^\top \mathbf{x} = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}^{-1}\mathbf{x} \geq \frac{1}{2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)^\top \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2) \quad (15)$$

com $\widehat{\mathbf{d}} = \mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)$.

Classificador de Bayes: exemplo 1

Exemplo 1. Suponha que $f_1(\mathbf{x})$ seja a densidade de uma distribuição normal padrão e $f_2(x)$ seja a densidade de uma distribuição normal com média 2 e variância 1. Supondo $\pi_1 = \pi_2 = 1/2$, elementos com valor das variáveis preditoras igual a $\mathbf{x}$ são classificados em $\mathcal{C}_1$ se $f_1(\mathbf{x})/f_2(\mathbf{x}) \geq 1$ o que equivale a

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = e^{-x^2/2} e^{(x-2)^2/2} \geq 1,$$

ou seja, se $x \leq 1$. Consequentemente, as duas probabilidades de classificação incorretas são iguais a $0,159$.

Classificador de Bayes: exemplo 2

Exemplo 2: Consideremos os dados do arquivo inibina, analisados por meio de regressão logística no Exemplo 6.9. Um dos objetivos é classificar as pacientes como tendo resposta positiva ou negativa ao tratamento com inibina com base na variável preditora `difinib = inibpos-inibpre`. Das 32 pacientes do conjunto de dados, $59,4\%$ apresentaram resposta positiva (classe C_1) e $40,5\%$ apresentaram resposta negativa (classe C_2).

Estimativas das médias das duas classes são, respectivamente, $\overline{x}_1 = 202,7$ e $\overline{x}_0 = 49,0$.
Estimativas das correspondentes variâncias são $S_1^2 = 31630,5$ e $S_0^2 = 2852,8$ e uma estimativa da variância comum é $S^2 = (18 \times S_1^2 + 12 \times S_0^2)/30 = 20119,4$.

De (11) obtemos o coeficiente da função discriminante
$d = (\overline{x}_1 - \overline{x}_0)/S^2 = 0,0076$. Para decidir em que classe uma paciente com valor de `difinib` $= x$ deve ser alocada, devemos comparar d com $(\overline{x}_1^2 - \overline{x}_0^2)/(2S^2) + [\log(\widehat{\pi}_0/\widehat{\pi}_1)] = 0,58191$.

Esses resultados podem ser concretizada por meio da função lda() do pacote MASS.

Classificador de Bayes: exemplo 2

```
lda(inibina$resposta ~ inibina$difinib, data = inibina)
Prior probabilities of groups:
negativa positiva
 0.40625  0.59375
Group means:
         inibina$difinib
negativa        49.01385
positiva       202.70158
Coefficients of linear discriminants:
                        LD1
inibina$difinib 0.007050054
```

A função considera as proporções de casos negativos (41%) e positivos (59%) no conjunto de dados de treinamento como probabilidades *a priori*, dado que elas não foram especificadas no comando.
O coeficiente da função discriminante (0.00705) corresponde à combinação linear de `difinib` usada para a decisão difere daquele obtido acima $(0,0076)$ pois a função `lda()` considera uma transformação com a finalidade de deixar os resultados com variância unitária (o que não influi na classificação).

Classificador de Bayes: exemplo 2

```
lda(inibina$resposta ~ inibina$difinib, data = inibina)
Prior probabilities of groups:
negativa positiva
 0.40625  0.59375
Group means:
         inibina$difinib
negativa        49.01385
positiva       202.70158
Coefficients of linear discriminants:
                        LD1
inibina$difinib 0.007050054
```

A função considera as proporções de casos negativos (41%) e positivos (59%) no conjunto de dados de treinamento como probabilidades *a priori*, dado que elas não foram especificadas no comando.
O coeficiente da função discriminante (0.00705) corresponde à combinação linear de `difinib` usada para a decisão difere daquele obtido acima $(0,0076)$ pois a função `lda()` considera uma transformação com a finalidade de deixar os resultados com variância unitária (o que não influi na classificação).

Classificador de Bayes: exemplo 2

Uma tabela relacionando a classificação predita com os valores reais da resposta pode ser obtido por meio dos comandos

```
predito <- predict(fisher)
table(predito$class, inibina$resposta)
           negativa positiva
  negativa         9        2
  positiva         4       17
```

indicando que a probabilidade de classificação correta é 81%, ligeiramente superior ao que foi conseguido com o emprego de regressão logística (ver Exemplo 6.9). Histogramas para os valores da função discriminante calculada para cada elemento do conjunto de dados estão dispostos na Figura 1.

Classificador de Bayes: exemplo 2

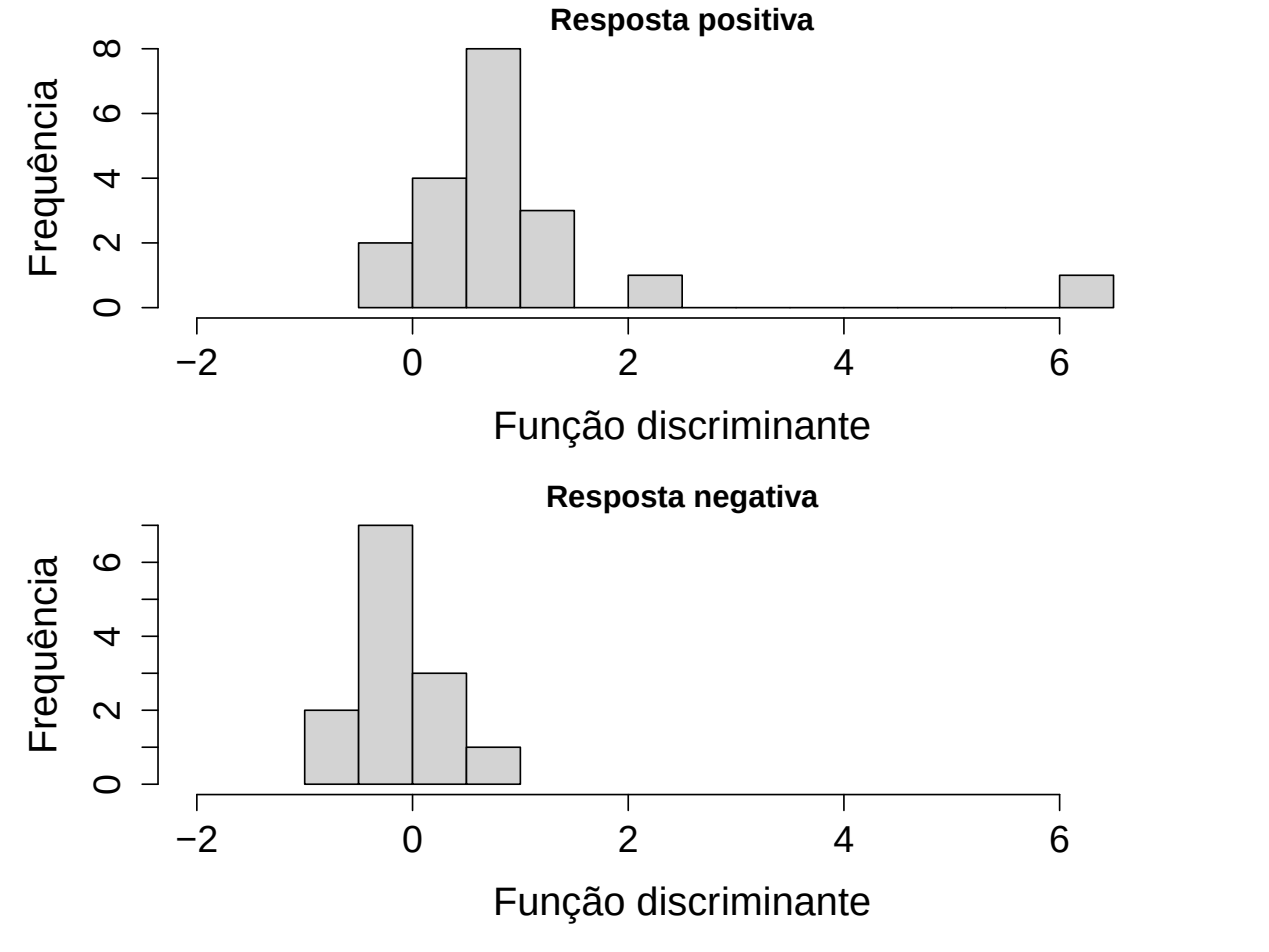

Figura 1: Histogramas para valores da função discriminante.

Função discriminante de Fisher

- Consideremos novamente o caso de duas classes (ou populações), $\mathcal{G}_1$ e $\mathcal{G}_2$ para as quais pretendemos obter um classificador com base em um vetor de variáveis preditoras, $\mathbf{X} = (X_1, \ldots, X_p)^\top$.
- A ideia de Fisher é considerar uma combinação linear $Y = \boldsymbol{\ell}^\top \mathbf{X}$, com $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_p)^\top$ de modo que o conjunto de variáveis preditoras seja transformado numa variável escalar Y.
- Sejam μ_{1Y} e μ_{2Y}, respectivamente, as médias de Y obtidas dos valores de $\mathbf{X}$ associadas aos dados $\mathcal{G}_1$ e $\mathcal{G}_2$. A regra para classificação consiste em selecionar a combinação linear que maximiza a distância quadrática entre essas duas médias, relativamente à variabilidade dos valores de Y.
- Uma suposição adicional e, às vezes, irrealista, é que as matrizes de covariâncias

$$\boldsymbol{\Sigma}_i = E(\mathbf{X} - \boldsymbol{\mu}_i)(\mathbf{X} - \boldsymbol{\mu}_i)^\top, \tag{16}$$

$i = 1, 2$, em que $\boldsymbol{\mu}_1 = E(\mathbf{X}|\mathcal{G}_1)$ e $\boldsymbol{\mu}_2 = E(\mathbf{X}|\mathcal{G}_2)$, sejam iguais para as duas classes, isto é, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.

FDLF

- Consequentemente,

$$\sigma_Y^2 = var(\ell^\top \mathbf{X}) = \ell^\top \mathbf{\Sigma} \ell$$

é igual para ambas as classes.

- $$\mu_{1Y} = E(Y|\mathcal{G}_1) = \ell^\top \boldsymbol{\mu}_1 \quad \text{e} \quad \mu_{2Y} = E(Y|\mathcal{G}_2) = \ell^\top \boldsymbol{\mu}_2$$

e a razão

$$\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2} = \frac{(\ell^\top \boldsymbol{\mu}_1 - \ell^\top \boldsymbol{\mu}_2)^2}{\ell^\top \mathbf{\Sigma} \ell} = \frac{\ell^\top (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \ell}{\ell^\top \mathbf{\Sigma} \ell}$$

$$= \frac{(\ell^\top \boldsymbol{\delta})^2}{\ell^\top \mathbf{\Sigma} \ell}, \tag{17}$$

com $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ é maximizada se

$$\ell = c\mathbf{\Sigma}^{-1}\boldsymbol{\delta} = c\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{18}$$

para todo $c \neq 0$.

- No caso $c = 1$, obtemos a função discriminante linear de Fisher

$$Y = \ell^\top \mathbf{X} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \mathbf{\Sigma}^{-1} \mathbf{X}. \tag{19}$$

e o valor máximo da razão (17) é $\boldsymbol{\delta}^\top \mathbf{\Sigma}^{-1} \boldsymbol{\delta}$.

FDLF

- Para uma nova observação $\mathbf{x}_0$, sejam $y_0 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x}_0$ e

$$\mu = \frac{\mu_{1Y} + \mu_{2Y}}{2} = \frac{1}{2}(\boldsymbol{\ell}^\top \boldsymbol{\mu}_1 + \boldsymbol{\ell}^\top \boldsymbol{\mu}_2) \tag{20}$$

(o ponto médio entre as médias univariadas associadas às duas classes). Em virtude de (19), esse ponto médio pode ser expresso como

$$\mu = \frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{2}, \tag{21}$$

- Consequentemente,

$$E(Y_0|\mathcal{G}_1) - \mu \geq 0 \quad \text{e} \quad E(Y_0|\mathcal{G}_2) - \mu < 0.$$

e uma regra de classificação é

Classifique $\mathbf{x}_0$ em $\mathcal{G}_1$ se $y_0 \geq \mu$,
Classifique $\mathbf{x}_0$ em $\mathcal{G}_2$ se $y_0 < \mu$.

Estimativa da FDLF

- Normalmente, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ e $\boldsymbol{\Sigma}$ são desconhecidas e têm que ser estimadas a partir de amostras de $\mathcal{G}_1$ e $\mathcal{G}_2$, denotadas por $\mathbf{X}_1 = [\mathbf{x}_{11}, \ldots, \mathbf{x}_{1,n_1}]$, uma matriz com dimensão $p \times n_1$ e $\mathbf{X}_2 = [\mathbf{x}_{21}, \ldots, \mathbf{x}_{n_2}]$, uma matriz com dimensão $p \times n_2$.
- Com os dados dessas amostras, podemos obter estimativas $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_1$ e $\mathbf{S}_2$, das médias e da matriz de covariâncias comum $\boldsymbol{\Sigma}$, para a qual um estimador não enviesado é

$$\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}. \tag{22}$$

- A função discriminante estimada é $\widehat{\ell}^\top \mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p \mathbf{x}$ e a regra de classificação é:

  Classifique a observação $\mathbf{x}_0$ em $\mathcal{G}_1$ se $y_0 - \widehat{\mu} \geq 0$,
  Classifique a observação $\mathbf{x}_0$ em $\mathcal{G}_2$ se $y_0 - \widehat{\mu} < 0$,

  em que $\widehat{\mu} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$.
- Nas demais expressões, os parâmetros são substituídas pelas respectivas estimativas.
- Outra suposição comumente adotada é que as variáveis preditoras têm distribuição Normal multivariada. Nesse caso, a solução encontrada por meio da função discriminante linear de Fisher é ótima.

Classificador KNN

- Vimos que o classificador de Bayes associa cada observação de teste com o valor do preditor x_0 à classe j de forma que

$$P(Y = j|X = x_0) \tag{23}$$

  seja a maior possível.
- No caso de duas classes, a observação será associada à Classe 1 se $P(Y = 1|X = x_0) > 0,5$ e à Classe 2, se $P(Y = 0|X = x_0) < 0,5$. A fronteira de Bayes é $P(Y = 1|X = x_0) = 0,5$.
- A taxa de erro de Bayes global é $1 - E(\max_j P(Y = j|X))$, obtida com base na média de todas as taxas de erro sobre todos os valores possíveis de j.
- Na prática como não conhecemos a distribuição condicional de Y, dado X, precisamos estimar essa probabilidade condicional, o que pode ser efetivado por meio de um método conhecido por K-ésimo vizinho mais próximo (*K-nearest neighbor*, KNN).

Classificador KNN

O algoritmo associado a esse método é:

i) Fixe K e uma observação teste x_0;

ii) Identifique K pontos do conjunto de dados de treinamento que sejam os mais próximos de x_0 segundo alguma medida de distância; denote esse conjunto por $\mathcal{V}_0$;

iii) Estime a probabilidade condicional de que a observação teste pertença à Classe j como a fração dos pontos de $\mathcal{V}_0$ cujos valores de Y sejam iguais a j, ou seja, como

$$P(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{V}_0} I(y_i = j). \tag{24}$$

iv) classifique x_0 na classe asociada à maior probabilidade.

A função knn() do pacote caret pode ser utilizada com essa finalidade.

Classificador KNN-Exemplo

- **Exemplo.** Consideremos, novamente, os dados do arquivo inibina utilizando a variável difinib como preditora e adotemos a estratégia de validação cruzada por meio do método LOOCV. Além disso, avaliemos o efeito de considerar entre 1 e 5 vizinhos mais próximos no processo de classificação.
- Os comandos necessários para a concretização da análise são

```
set.seed(2327854)
trControl <- trainControl(method  = "LOOCV")

fit <- train(resposta ~ difinib, method = "knn",
             tuneGrid   = expand.grid(k = 1:5),
             trControl  = trControl, metric= "Accuracy",
             data = inibina)
fit
```

Classificador KNN-Exemplo

Os resultados correspondentes são:

```
k-Nearest Neighbors

32 samples
 1 predictor
 2 classes: 'negativa', 'positiva'

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 31, 31, 31, 31, 31, 31, ...
Resampling results across tuning parameters:

  k  Accuracy  Kappa
  1  0.71875   0.4240000
  2  0.78125   0.5409836
  3  0.81250   0.6016598
  4  0.78125   0.5409836
  5  0.81250   0.6016598

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 5.
```

Classificador KNN-Exemplo

- Segundo o esse processo, o melhor resultado (com K=5 vizinhos) gera uma acurácia (média) de 81.3%. A tabela de classificação obtida por meio do ajuste do modelo final ao conjunto de dados original, juntamente com estatísticas descritivas pode ser obtida por meio dos comandos:

```
predito <- predict(fit)
confusionMatrix(predito, inibina\$resposta)
```

- que geram os seguintes resultados:

```
Confusion Matrix and Statistics
          Reference
Prediction negativa positiva
  negativa        9        1
  positiva        4       18

               Accuracy : 0.8438
                 95% CI : (0.6721, 0.9472)
    No Information Rate : 0.5938
    P-Value [Acc > NIR] : 0.002273
                  Kappa : 0.6639
 Mcnemar's Test P-Value : 0.371093
```

Classificador KNN-Exemplo

```
            Sensitivity : 0.6923
            Specificity : 0.9474
         Pos Pred Value : 0.9000
         Neg Pred Value : 0.8182
             Prevalence : 0.4062
         Detection Rate : 0.2812
   Detection Prevalence : 0.3125
      Balanced Accuracy : 0.8198
```

A acurácia é de $84,4\%$, sensibilidade de $69,2\%$ e especificidade de $94,7\%$.

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= $\frac{TPR+TNR}{2}$

Teste de McNemar

- É um teste para dados nominais pareados, dispostos numa tabela de contingência 2×2 (McNemar, 1947).
- Considere a tabela, com resultados de dois testes para n indivíduos:

		Teste 2		
		Positivo	Negativo	
Teste 1	Positivo	n_{11}	n_{12}	$n_{1\cdot}$
Teste 1	Negativo	n_{21}	n_{22}	$n_{2\cdot}$
		$n_{\cdot 1}$	$n_{\cdot 2}$	1

- Sejam $p_{ij} = n_{ij}/n$, $i, j = 1, 2$, $p_{i\cdot}$ as soma das linhas, $i = 1, 2$ $p_{\cdot j}$ as somas das colunas, $j = 1, 2$ (com os p_{ij} substituindo os n_{ij} na tabela). A hipótese nula de homogeneidade marginal afirma que as duas probabilidades marginais de cada resultado são iguais, isto é, $p_{11} + p_{12} = p_{11} + p_{21}$ e $p_{21} + p_{22} = p_{12} + p_{22}$, ou seja, $p_{1\cdot} = p_{\cdot 1}$ e $p_{2\cdot} = p_{\cdot 2}$, ou ainda

$$H_0 : p_{12} = p_{21},$$

$$H_1 : p_{12} \neq p_{21}.$$

- A estatística de MacNemar é

$$M = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}},$$

que, sob H_0, tem uma distribuição qui-quadrado com 1 grau de liberdade.

Teste de McNemar

- Se n_{12} ou n_{21} for pequeno (soma<25), então M não é bem aproximada pela distribuição qui-quadrado. Um teste binomial exato pode ser usado para n_{21}. Para $n_{12} > n_{21}$, o valor-p exato é

$$p = 2 \sum_{i=n_{12}}^{N} \binom{N}{i} (1/2)^i (1/2)^{N-i},$$

  com $N = n_{12} + n_{21}$.
- Edwards (1948) propôs a seguinte correção de continuidade para M:

$$M = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}.$$

- No exemplo, $M = (|1-4|-1)^2/5 = 0,8$ e valor-p é $P(\chi^2_1 > 0,8|H_0) \approx 0,37$, logo não rejeitamos H_0.

Aplicação do teste em ML

- O teste de McNemar pode ser usado para comparar técnicas de classificação, para aqueles algoritmos que são usados em conjuntos de dados grandes e não podem ser repetidos via algum método de reamostragem, como CV.
- Dietterich (1998) considerou 5 testes para determinar se um algoritmo de classificação é melhor do que um outro, em um particular conjunto de dados. O objetivo era determinar a probabilidade de erro de tipo I de cada teste.
- Testes que não devem ser usados:(a) teste para a diferença de duas proporções; (b) teste t pareado (diferenças) baseado em partições aleatórias dos conjuntos de treinamento/teste; (c) teste t pareado baseado em 10-fold CV. Todos exibem probabilidades de erro de tipo I altas.
- O teste de McNemar tem baixa probabilidade de erro do tipo I.
- O autor introduziu um teste, 5×2 CV, baseado em 5 iterações de uma 2-fold CV, que tem uma probabilidade de erro de tipo I aceitável e poder maior do que o teste de McNemar.

Aplicação do teste em ML

- Questão: Dados dois classificadores C_1 e C_2 e dados suficientes para aplicá-los em um conjunto de teste, determinar qual classificador será mais acurado em novos conjuntos de testes.
- Essa questão pode ser respondida medindo-se a acurácia de cada classificador no conjuto teste a aplicando o teste de McNemar.
- Dietterich (1998) considera 9 questões , algumas ainda não respondidas, e foca seu artigo na seguinte questão: Dados dois algoritmos de aprendizagem A e B, e um conjuntos de dados pequeno S, qual algoritmo produzirá classificadores mais acurados quando treinados em conjuntos de dados do mesmo tamanho que S?
- Para isso, é necessário usar métodos de reamostragem. Ele compara vários testes estatísticos para responder a questão.
- O primeiro passo é identificar as fontes de variação que podem ser controladas por cada teste.
- 4 fontes de variação: (a) variação aleatória na seleção do conjunto de teste que será usado para avaliar os algoritmos; (b) variação na escolha dos dados de treinamento (instabilidade); (c) aleatoriedade interna do algoritmo de aprendizagem. Por exemplo, o algoritmo backpropagation depende dos pesos (aleatórios) iniciais; (d) erro de classificação aleatório.

Aplicação do teste em ML

- Um teste deve concluir que dois algoritmos são diferentes se, e somente se, suas taxas de classificação corretas sejam diferentes, em média, quando treinados em um conjunto de treinamento de tamanho fixo e testado em todos os dados da população.
- Para tanto, o teste deve considerar o tamanho do teste (probabilidade do erro de tipo I) e excutar o algoritmo múltiplas vezes e medir a variação da acurácia dos classificadores resultantes
- Para aplicar o teste de McNemar, dividimos a amostra de dados S em um conjunto de treinamento T_0 (com n observações) e um conjunto teste T_1 (com m observações). Treinamos os algoritmos C_1 e C_2 no conjunto T_0, obtendo-se classificadores $\hat{C}_1$ e $\hat{C}_2$. Então, testamos esses classificadores no conjunto T_1. Para cada $\mathbf{x} \in T_1$, registramos como esse ponto foi classificado e construímos a seguinte tabela 2×2:

		Classificador 2	
		Class. Correta	Class. Errônea
Classificador 1	Class. correta	n_{11} = Sim/Sim	n_{12}=Sim/Não
Classificador 1	Class. errôna	n_{21} =Não/Sim	n_{22}=Não/Não

- $\sum_i \sum_j n_{ij} = m$.

Aplicação do teste em ML

- Rejeitandos-e H_0, os dois algoritmos terão desempenho diferentes quando treinados em T_0.
- Note que esse teste tem dois problemas: primeiro, não mede diretamente a variabilidade devida à escolha de T_0, nem a aleatoriedade interna do algoritmo, pois um único conjunto de treinamento é escolhido. Segundo, ele não compara os desempenhos dos algoritmos em conjuntos de treinamento de tamanho $|S|$, mas sobre conjuntos de tamanho n, que deve ser menor do que $|S|$, para que tenhamos um conjunto de teste grande.

AD quadrática

- O classificador obtido por meio de Análise Discriminante Quadrática supõe, como na Análise Discriminante Linear usual, que as observações são extraídas de uma distribuição gaussiana multivariada, mas não necessita da suposição de homocedasticidade (matrizes de covariâncias iguais), ou seja, admite que a cada classe esteja associada uma matriz de covariância, $\mathbf{\Sigma}_k$.
- Como no caso da Análise Discriminante Linear, não é difícil ver que o classificador de Bayes associa um elemento com valor das variáveis preditora igual a $\mathbf{x}$ à classe para a qual a função quadrática

$$\begin{aligned} \delta_k(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log \pi_k \qquad (25) \\ &= -\frac{1}{2}\mathbf{x}^\top \mathbf{\Sigma}_k^{-1}\mathbf{x} + \mathbf{x}^\top \mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^\top \mathbf{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log \pi_K. \end{aligned}$$

é **máxima**.

AD quadrática

- Como os elementos de (25) não são conhecidos, pode-se estimá-los por meio de

$$\bar{\mathbf{x}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}\mathbf{x}_i, \quad \mathbf{S}_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^\top$$

  e $\widehat{\pi}_k = n_k/n$, em que n_k é o número de observações na classe k e n é o número total de observações no conjunto de dados. O primeiro termo do segundo membro da primeira igualdade de (25) é a distância de Mahalanobis.
- O número de parâmetros a estimar, $Kp(p+1)/2$, é maior que no caso de Análise Discriminante Linear, na qual a matriz de covariâncias é comum. Além disso, a versão linear apresenta variância substancialmente menor mas viés maior do que a versão quadrática. A Análise Discriminante Quadrática é recomendada se o número de dados for grande; em caso contrário, convém usar Análise Discriminante Linear.

AD regularizada

- A Análise Discriminante Regularizada foi proposta por Friedman (1989) e é um compromisso entre Análise Discriminante Linear e Análise Discriminante Quadrática.
- O método proposto por Friedman consiste em "encolher" (*shrink*) as matrizes de covariâncias da Análise Discriminante Quadrática em direção a uma matriz de covariâncias comum.
- Friedman (1989) propõe o seguinte procedimento de regularização

$$\boldsymbol{\Sigma}_k(\lambda, \gamma) = (1 - \gamma)\boldsymbol{\Sigma}_k(\lambda) + \frac{\gamma}{p}\text{tr}[\boldsymbol{\Sigma}_k(\lambda)]\mathbf{I}, \tag{26}$$

em que $\text{tr}(\mathbf{A})$ indica o traço da matriz $\mathbf{A}$, $\mathbf{I}$ é a matriz identidade e

$$\boldsymbol{\Sigma}_k(\lambda) = \lambda\boldsymbol{\Sigma}_k + (1 - \lambda)\boldsymbol{\Sigma} \tag{27}$$

com $\boldsymbol{\Sigma} = \sum n_k\boldsymbol{\Sigma}_k/n$.

AD regularizada

- O parâmetro $\lambda \in [0, 1]$ controla o grau segundo o qual a matriz de covariâncias ponderada pode ser usada e $\gamma \in [0, 1]$ controla o grau de encolhimento ao autovalor médio. Na prática, λ e γ são escolhidos por meio de LOOVC para cada ponto de uma grade no quadrado unitário.
- Quando $\lambda = 1$ e $\gamma = 0$, a Análise Discriminante Regularizada reduz-se à Análise Discriminante Linear. Se $\lambda = 0$ e $\gamma = 0$, o método reduz-se à Análise Discriminante Quadrática. Se $p > n_k$, para todo k e $p < n$, $\mathbf{\Sigma}_k$ é singular, mas nem a matriz ponderada $\mathbf{\Sigma}$ nem $\mathbf{\Sigma}_k(\lambda)$ em (27) o são. Se $p > n$, todas essas matrizes são singulares e a matriz $\mathbf{\Sigma}_k(\lambda, \gamma)$ é regularizada por (26).
- Essas análises podem ser concretizadas por meio das funções `qda()` do pacote `MASS` e `rda()` do pacote `klaR`.

ADQ e ADR - exemplo

Vamos considerar novamente o conjunto de dados `disco` do Exemplo 8.1. Na segunda análise realizada por meio de regressão logística, a classificação foi concretizada via validação cruzada VC5/5 e LOOCV tendo como variáveis preditoras a distância aberta, a distância fechada ou ambas.
A melhor acurácia, $85,7\%$ foi obtida com ambas as variáveis preditoras e VC5/5.
Agora, consideramos a classificação realizada por intermédio de Análises Discriminantes Linear, Quadrática e Regularizada, separando os dados em um conjunto de treinamento contendo 80% (83) dos elementos, selecionados aleatoriamente, e em um conjunto de validação com os restantes 21 elementos.
Lembremos que $y = 1$ corresponde a discos deslocados e $y = 0$ a discos não deslocados.
As acurácias obtidas por meio de Análise Discriminante Linear e Análise Discriminante Quadrática foram, respectivamente, 90% e 85%.

Referências

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, **10**, 1895–1923.

Edwards, A (1948). Note on the correction for continuity in testing the significance of the difference between correlated proportions. *Psychometrika*, **13**, 185–187.

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

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12**, 153–157.

Morettin, P. A. e Singer, J. M. (2022). *Estatística e Ciência de Dados*. LTC: Rio de Janeiro.