Análise de Discriminantes e Classificação

Prof. George von Borries

Departamento de Estatística
Universidade de Brasília

2023

Introdução

Discriminantes:

Determinar as regras que indicam as diferenças entre objetos provenientes de diferentes populações conhecidas (Fisher, 1938).

Classificação:

Alocar novos objetos em duas ou mais classes (populações) de acordo com regras estabelecidas e avaliar a qualidade da alocação.

Os dois conceitos se confundem e na realidade constituem uma única atividade.

Literatura:

Aprendizado Supervisionado, Aprendizado Estatístico, Aprendizado de Máquina, Reconhecimento de Padrões, Mineração de Dados.

Exemplos de Aplicações

Reconhecimento de Faces, identificação de fraudes, análise de qualidade, sistemas de segurança, reconhecimento de voz, sites de busca, estímulos via sinais biopontenciais.

Aprendizado Estatístico em Classificação

Consiste de três fases principais:

- Extração de características

  Característica: quaisquer valores, estatísticas, padrões ou variáveis utilizadas para discriminar duas ou mais classes.

  Extração de características: quaisquer técnicas ou transformações que, aplicadas a um determinado conjunto de dados, produzem um outro conjunto formado por características.

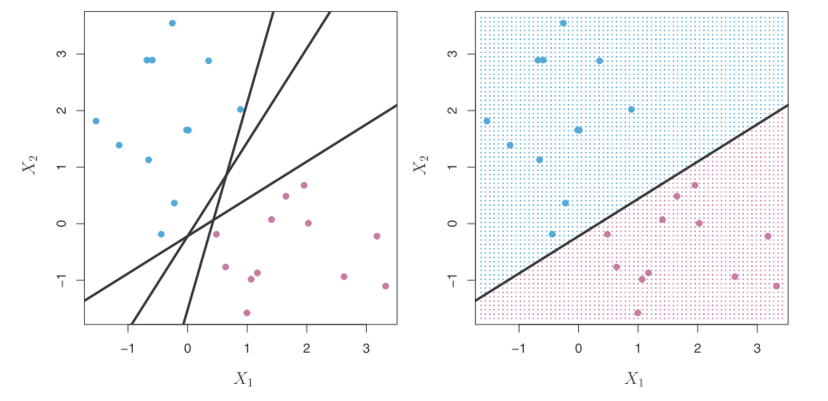
- Treinamento da função discriminante: ajuste da função com base em alguma métrica de erro em relação a variável de supervisão. O ajuste é feito em dados de treinamento.

- Teste ou classificação: Para um conjunto de dados em separado, verificar a qualidade da função discriminante ajustada.

Nota: Em geral, a extração e treinamento são denominadas de análise de discriminantes.

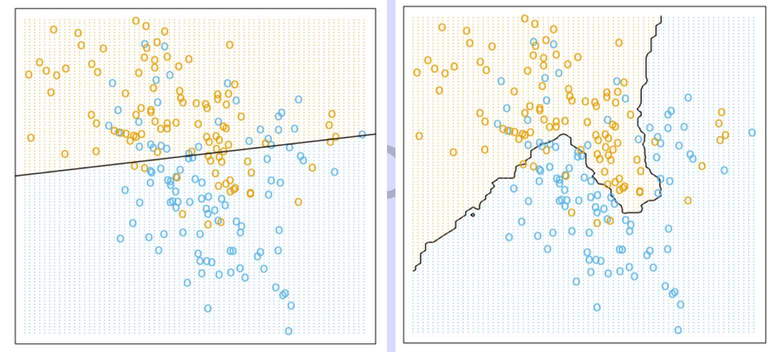
Objetivo

Encontrar funções que melhor separem grupos ou populações.



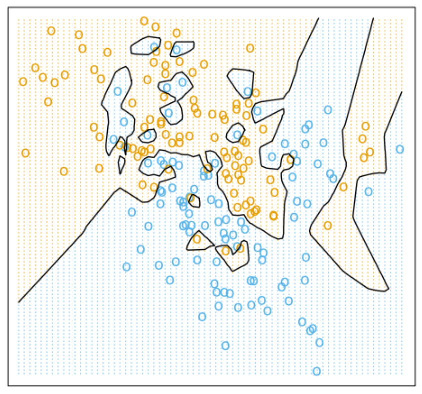
Fonte: James, Witten, Hastie, Tibshirani, 2017.

As funções discriminantes podem ser lineares ou não lineares.



Fonte: Hastie, Tibshirani e Friedman, 2009.

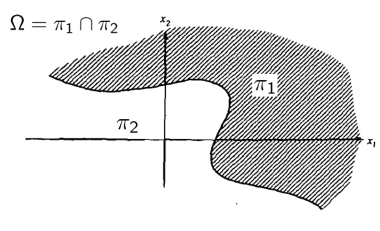
E em alguns casos os discriminantes podem ser bem complicados.



Fonte: Hastie, Tibshirani e Friedman, 2009.

Discriminante para duas populações

- Seja Ω o espaço amostral formado pelas regiões (classes, populações) π_1 e π_2 tal que $\Omega = \pi_1 \cap \pi_2$.



- $\mathbf{X} = [\mathrm{X}_1 \;\; \mathrm{X}_2 \;\; \ldots \;\; \mathrm{X}_p]$ é formado por p características que descrevem as duas populações, tal que

  π_1: população 1 que pode ser descrita por $f_1(\mathbf{X})$.

  π_2: população 2 que pode ser descrita por $f_2(\mathbf{X})$.

  ($f(\cdot)$ é a função densidade de probabilidade)

Discriminante para duas populações

Regra de Classificação de Bayes

- Seja $\mathrm{P}(\mathbf{X} \in \pi_i) = p_i, \quad i = 1, 2$ a probabilidade a priori que uma observação $\mathbf{X} = \mathbf{x}$ pertence a população π_1 ou π_2.
- $f_i(\mathbf{x}) = \mathrm{P}(\mathbf{X} = \mathbf{x} | \mathbf{X} \in \pi_i), \quad i = 1, 2.$ é a densidade de probabilidade condicional multivariada de $\mathbf{X}$ para a i-ésima população ou classe.
- Pelo Teorema de Bayes,

$$\mathrm{P}(\pi_i|\mathbf{x}) = \mathrm{P}(\mathbf{X} \in \pi_i | \mathbf{X} = \mathbf{x}) = \frac{f_i(\mathbf{x}) p_i}{f_1(\mathbf{x}) p_1 + f_2(\mathbf{x}) p_2}$$

é a probabilidade posterior de que o valor observado $\mathbf{x}$ pertence a $\pi_i, \quad i = 1, 2.$

Discriminante para duas populações

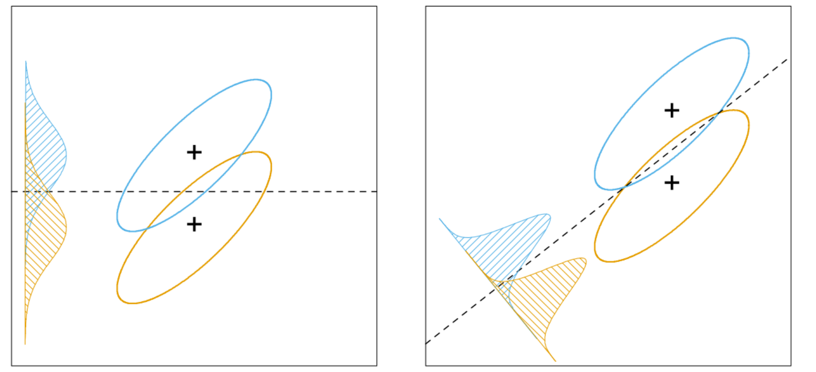
Regra de Classificação de Bayes

- O Classificador de Bayes aloca **x** a classe de maior probabilidade posterior, i.e., se
$$\frac{\mathrm{P}(\pi_1|\mathbf{x})}{\mathrm{P}(\pi_2|\mathbf{x})} \geq 1,$$
alocar **x** a π_1 e alocar e π_2 caso contrário.

- Utilizando o Terema de Bayes, a mesma regra pode ser escrita como
$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1}.$$

- No caso de igualdade da razão do Classificador de Bayes a alocação pode ser feita por sorteio.

Discriminante Linear para duas Populações N_p



Separações obtidas por funções discriminantes (Hastie et al., 2009).

Discriminante Linear para duas Populações N_p

Caso 1: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (Suposição de Homogeneidade)

Seja $\mathbf{X} = [\mathrm{X}_1\ \mathrm{X}_2\ \ldots\ \mathrm{X}_p]$ e

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \quad i = 1, 2.$$

Então

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}}$$

e

$$\begin{aligned}\ell(\mathbf{x}) = \ln\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= \overbrace{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}^{\text{Disc. Linear de Fisher}} - \frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),\end{aligned}$$

em que $\boldsymbol{\mu} = (\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)/2$.

$\Rightarrow$ Se $\ell(\mathbf{x}) > 0$ alocar $\mathbf{x}$ a π_1.

Observação: Note que $\ell(\mathbf{x})$ pode ser escrito como uma função linear $\beta_0 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}$.

Discriminante Linear para duas Populações N_p

Caso 1: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ (Suposição de Homogeneidade)

Na prática substituimos $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ por $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \mathbf{S}_{\mathrm{pon}}$,

$$\mathbf{S}_{\mathrm{pon}} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2$$

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

e alocamos $\boldsymbol{x}$ a π_1 se

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\mathrm{pon}}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\mathrm{pon}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) > 0.$$

Caso contrário alocar $\mathbf{x}$ a π_2.

Discriminante Quadrática para duas Populações N_p

Caso 2: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Seja $\mathbf{X} = [\mathrm{X}_1\ \mathrm{X}_2\ \ldots\ \mathrm{X}_p]$ e

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathsf{T}}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \quad i = 1, 2.$$

Então

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{(2\pi)^{p/2}|\boldsymbol{\Sigma}_2|^{1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right\}}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_1|^{1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\}}$$

e

$$\begin{aligned}
\ell(\mathbf{x}) = \ln\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} &= c_0 - \frac{1}{2}\left\{(\mathbf{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)\right\} \\
&= c_1 - \frac{1}{2}\mathbf{x}^{\mathsf{T}}\left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right)\mathbf{x} + \left(\boldsymbol{\mu}_1^{\mathsf{T}}\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^{\mathsf{T}}\boldsymbol{\Sigma}_2^{-1}\right)\mathbf{x}
\end{aligned}$$

em que c_0 e c_1 são constantes que dependem somente dos parâmetros $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$.

$\Rightarrow$ Se $\ell(\mathbf{x}) > 0$ alocar $\mathbf{x}$ a π_1.

Discriminante Quadrática para duas Populações N_p

Caso 2: $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Observação:

Note que $\ell(\mathbf{x})$ pode ser escrito como uma função linear

$$\mathrm{Q}(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^\top \mathbf{x} + \mathbf{x}^\top \mathbf{w} \mathbf{x},$$

sendo

$$\mathbf{w} = \frac{1}{2}\left(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}\right), \qquad \boldsymbol{\beta} = \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2$$

e

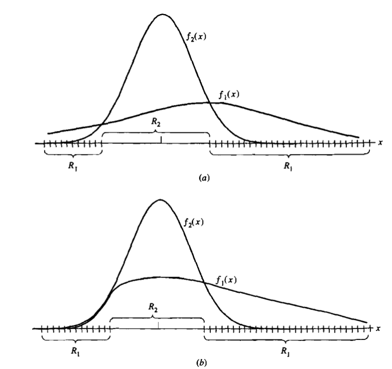
$$\beta_0 = -\frac{1}{2}\left\{\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} + \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^\top\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right\} - \ln\left(\frac{p_2}{p_1}\right)$$

$\mathrm{Q}(\mathbf{x})$ é denominado função discriminante quadrática.

Discriminante Quadrática para duas Populações N_p

Caso 2: $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$

Regra Quadrática (não assume homogeneidade das matrizes de variância-covariância) quando (a) duas normais, (b) uma normal e uma não normal, resultando numa regra inadequada.



(Fonte: Johnson & Wichern, Fig. 11.6)

Soluções:

- Transformação para normalizar.
- Usar parte dos dados para treinamento e parte para validação.

Qualidade da Regra de Discriminação

Teste de Hotelling

Assumindo $\mathbf{X}_1 \sim \mathrm{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ e $\mathbf{X}_2 \sim \mathrm{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. A estatística do teste

$\mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ vs
$\mathrm{H}_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

será

$$\mathrm{F} = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} \mathrm{T}^2 \sim \mathrm{F}_{p, n_1 + n_2 - p - 1}$$

com

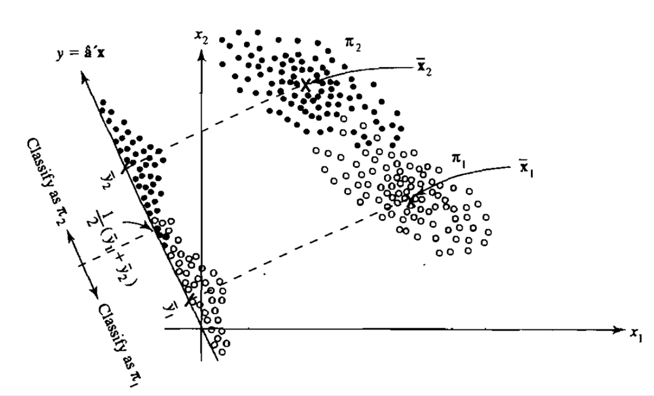
$$\mathrm{T}^2 = \frac{n_1 n_2}{n_1 + n_2} \mathrm{D}^2$$

e $\mathrm{D}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$.

Se a diferença entre a média dos escores das duas populações não for significativa, a função discriminante deve ser reformulada com o objetivo de buscar variáveis com maior poder de discriminação dos grupos.

Função Discriminante de Fisher (duas Populações)

- A ideia é produzir observações univariadas y de π_1 e π_2 que possuem a maior separação possível.
- A técnica não assume normalidade das populações, mas assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.



Representação do Discriminante de Fisher com $p = 2$. Fonte: J&W, Fig. 11.5.

Função Discriminante de Fisher (duas Populações)

- Procuramos funções lineares ($\hat{\mathbf{a}}^\top \mathbf{x}$), que maximizam a separação entre $\bar{\mathrm{x}}_1$ e $\bar{\mathrm{x}}_2$ dos dados observados.
- Resultado

  A combinação linear $\hat{\mathrm{x}} = \hat{\mathbf{a}}^\top \mathbf{x} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\mathrm{pon}}^{-1} \mathbf{x}$ maximiza a razão

  $$\frac{(\hat{\mathbf{a}}^\top \mathbf{x}_1 - \hat{\mathbf{a}}^\top \mathbf{x}_2)^2}{\hat{\mathbf{a}}^\top \mathbf{S}_{\mathrm{pon}} \hat{\mathbf{a}}} = \frac{(\mathbf{a}^\top \mathbf{d})^2}{\hat{\mathbf{a}}^\top \mathbf{S}_{\mathrm{pon}} \hat{\mathbf{a}}}$$

  em relação a todos os possíveis coeficientes de $\hat{\mathbf{a}}$.
- A distância de Mahalanobis $\mathrm{D}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\mathrm{pon}}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ é o máximo da razão.
- Note que $(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_{\mathrm{pon}}^{-1} \mathbf{x}$ é o Discriminante Linear de Fisher para os dados observados.
- Sejam $\hat{y} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1} \mathbf{x}$ e $\hat{\mathrm{m}} = \frac{1}{2} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^\top \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)$. Alocar $\mathbf{x}$ a π_1 se $\hat{y} \geq \hat{\mathrm{m}}$. Caso contrário, alocar $\mathbf{x}$ a π_2.

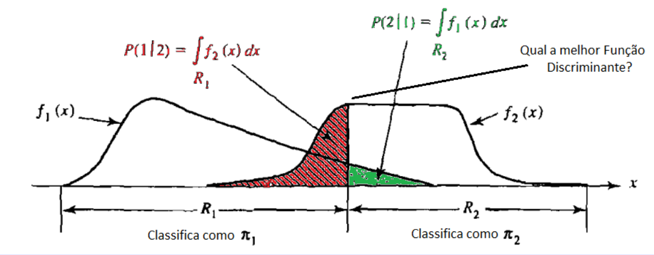
Erro de Classificação com Duas Populações

- Seja R_i a região amostral que determina a alocação de $\mathbf{x}$ em π_i.
- Seja $\mathrm{P}(2|1)$ a probabilidade condicional de classificação de um objeto em π_2 quando o objeto pertence a π_1, i.e.,

$$\mathrm{P}(2|1) = \mathrm{P}(\mathbf{X} \in R_2|\pi_1) = \int_{R_2} f_1(\mathbf{x})d\mathbf{x}$$

- Da mesma forma podemos definir $P(1|2)$,

$$\mathrm{P}(1|2) = \mathrm{P}(\mathbf{X} \in R_1|\pi_2) = \int_{R_1} f_2(\mathbf{x})d\mathbf{x}$$



Fonte: Johnson & Wichern, Fig. 11.3.

Erro de Classificação com Duas Populações

A probabilidade de **classificação correta** é dada por,

$$\mathrm{P}(\mathbf{X} \in \mathrm{R}_1 \cap \mathbf{X} \in \pi_1) = \mathrm{P}(\mathbf{X} \in \mathrm{R}_1 | \mathbf{X} \in \pi_1)\, \mathrm{P}(\mathbf{X} \in \pi_1) = \mathrm{P}(1|1) p_1,$$

$$\mathrm{P}(\mathbf{X} \in \mathrm{R}_2 \cap \mathbf{X} \in \pi_2) = \mathrm{P}(\mathbf{X} \in \mathrm{R}_2 | \mathbf{X} \in \pi_2)\, \mathrm{P}(\mathbf{X} \in \pi_2) = \mathrm{P}(2|2) p_2$$

A probabilidade de **classificação incorreta** é obtida por,

$$\mathrm{P}(\mathbf{X} \in \mathrm{R}_1 \cap \mathbf{X} \in \pi_2) = \mathrm{P}(\mathbf{X} \in \mathrm{R}_1 | \mathbf{X} \in \pi_2)\, \mathrm{P}(\mathbf{X} \in \pi_2) = \mathrm{P}(1|2) p_2,$$

$$\mathrm{P}(\mathbf{X} \in \mathrm{R}_2 \cap \mathbf{X} \in \pi_1) = \mathrm{P}(\mathbf{X} \in \mathrm{R}_2 | \mathbf{X} \in \pi_1)\, \mathrm{P}(\mathbf{X} \in \pi_1) = \mathrm{P}(2|1) p_1$$

Custos de erro de classificação podem ser incluídos no estudo, sendo

$c(1|2)$ o custo de classificar em π_1 um elemento de π_2.
$c(2|1)$ o custo de classificar em π_2 um elemento de π_1.

Custo Esperado do Erro de Classificação (ECM)

- O objetivo é definir R_1 e R_2 que minimizam

$$\mathrm{ECM} = c(2|1)\mathrm{P}(2|1)p_1 + c(1|2)\mathrm{P}(1|2)p_2$$

(*Expected Cost of Missclassification*)

- Estas regiões correspondem a regra definida pelo Teorema de Bayes, considerando agora os custos de classificação, i.e., se

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

classificar $\mathbf{x}$ em π_1. Caso contrário, classificar em π_2.

Probabilidade Total do Erro de Classificação (TPM)

- $$\mathrm{TPM} = p_1 \int_{\mathrm{R}_2} f_1(\mathbf{x})d\mathbf{x} + p_2 \int_{\mathrm{R}_1} f_2(\mathbf{x})d\mathbf{x}$$

(*Total Probability of Missclassification*)

$\mathrm{ECM} = \mathrm{TPM}$ se $c(2|1) = c(1|2)$.

Avaliação das Funções de Classificação

Três conceitos aparecem na avaliação de funções de classificação:

OER (*Optimum Error Rate*) taxa de erro que minimiza TPM. O problema é que precisamos de $f_1, f_2, \mathrm{R}_1, \mathrm{R}_2$ em TPM.

AER (*Actual Error Rate*) obtem $\hat{\mathrm{R}}_1, \hat{\mathrm{R}}_2$ através de amostras de treinamento, indicando como a função irá atuar em amostras futuras. O problema é que ainda precisa de f_1, f_2 em TPM.

APER (*Apparent Error Rate*) indica a fração de observações na amostra de treinamento que são incorretamente classificadas. Neste caso as amostras de cada população devem ser relativamente grandes, senão irá subestimar TPM.

O processo de estimação de APER é feito por validação cruzada Jackknife.

Avaliação das Funções de Classificação

Estimação de APER por validação cruzada Jacknife. Considere n observações na amostra.

1. Faça $i = 0$.
2. Faça $i = i + 1$, retire a observação i do conjunto de dados e retorne a observação $i - 1$.
3. Obtenha a função de classificação com os dados restantes.
4. Classifique a observação i utilizando a função em **3**.
5. Se $i < n$ retornar a **2.**. Se $i = n$, estimar APER por

$$\widehat{\mathrm{APER}} = \frac{e_1 + e_2}{n}$$

em que e_i é número de observações incorretamente classificadas na população π_i em **4**.

Função Discriminante de Fisher para $g > 2$ Populações

- A técnica considera $g > 2$ populações e assume homogeneidade, i.e., $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \ldots = \mathbf{\Sigma}_g = \mathbf{\Sigma}$.
- Assim como no caso binário, procura-se funções lineares (discriminantes de Fisher) que melhor separam os grupos.
- Seja

$$\mathrm{SQEG} = \sum_{i=1}^{g} n_i \mathbf{u}^\top (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})^\top (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}) \mathbf{u} = \mathbf{u}^\top \mathbf{B} \mathbf{u} \quad \text{(variação entre grupos)}$$

e

$$\mathrm{SQDG} = \sum_{i=1}^{g} (n_i - 1) \mathbf{u}^\top \mathbf{S}_i \mathbf{u} == \sum_{i=1}^{g} \sum_{j=1}^{n_i} \mathbf{u}^\top (\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_i)^\top (\bar{\mathbf{x}}_{ij} - \bar{\mathbf{x}}_i) \mathbf{u}$$

$$= \mathbf{u}^\top \mathbf{W} \mathbf{u} \quad \text{(variação dentro de grupos)}$$

Função Discriminante de Fisher para $g > 2$ Populações

- Para que uma função f seja discriminante, ela deve variar muito entre grupos e pouco dentro de grupos, i.e., devemos obter $\mathbf{u}$ de forma a maximizar

$$\mathrm{b} = \frac{\mathrm{SQEG}}{\mathrm{SQDG}} = \frac{\mathbf{u}^\top \mathbf{B} \mathbf{u}}{\mathbf{u}^\top \mathbf{W} \mathbf{u}}$$

- O problema é resolvido derivando b em relação a $\mathbf{u}$ e igualando a zero. A solução é obtida por,

$$(\mathbf{W}^{-1}\mathbf{B})\mathbf{u} = \mathrm{b}\mathbf{u}$$

ou seja, b é um autovalor de $\mathbf{W}^{-1}\mathbf{B}$ e $\mathbf{u}$ seu respectivo autovetor.

- Assim, $f = \mathbf{u}^\top \mathbf{x}$ com $\mathbf{u}$ o autovetor associado ao maior autovalor de $\mathbf{W}^{-1}\mathbf{B}$.

Função Discriminante de Fisher para $g > 2$ Populações

- Sejam $\lambda_1 > \cdots > \lambda_g > 0$ os autovalores não nulos de $\mathbf{W}^{-1}\mathbf{B}$ e $\mathbf{e}_1, \ldots, \mathbf{e}_g$ os correspondentes autovetores, tais que $\mathbf{e}^\top \mathbf{S}_{\text{pon}} \mathbf{e} = 1$. O vetor de coeficientes $\mathbf{u}$ que maximiza a razão,

$$\mathrm{b} = \frac{\mathbf{u}^\top \mathbf{B} \mathbf{u}}{\mathbf{u}^\top \mathbf{W} \mathbf{u}}$$

obtida por $\mathbf{u}_1 = \mathbf{e}_1$.

Assim, falamos em

$\mathbf{u}_1\mathbf{x}$ o primeiro discriminante amostral.
$\mathbf{u}_2\mathbf{x}$ o segundo discriminante amostral.
$\vdots$
$\mathbf{u}_g\mathbf{x}$ o g-ésimo discriminante amostral.

Função Discriminante de Fisher para $g > 2$ Populações

- Regra de Separação: Alocar $\mathbf{x}$ a π_g se,

$$\sum_{j=1}^{r} \left[\hat{\mathbf{a}}_j^T(\mathbf{x} - \bar{\mathbf{x}}_g)\right] \leq \sum_{j=1}^{r} \left[\hat{\mathbf{a}}_j^T(\mathbf{x} - \bar{\mathbf{x}}_i)\right] \quad \text{para todo} \quad i \neq g.$$

- Ver Exemplo 11.16 de Johnson e Wichern, pág. 631 a 633.

- O poder de separação é dado pelas seguintes estatísticas:
  1. Lambda de Wilks: $|\mathbf{W}|/|(\mathbf{B} + \mathbf{W})|$
  2. Traço de Pillai: $\mathrm{tr}(\mathbf{B}(\mathbf{B} + \mathbf{W}^{-1}))$
  3. Traço de Hotelling-Lawley: $\mathrm{tr}(\mathbf{W}^{-1}\mathbf{B})$
  4. Raiz máxima de Roy: maior autovalor de $\mathbf{W}^{-1}\mathbf{B}$

  Os testes utilizam a distribuição F com respectivos graus de liberdade.

Exemplos

1. Johnson e Wichern, Exemplo 11.8 - Dados de Salmão

   - Pesca de salmão é recurso valioso para EUA e Canadá.

   - Existe a preocupação de que pescadores do Alaska pesquem muito salmão do Canadá e vice-versa.

   - O Salmão nasce em rios de água doce e após um ou dois anos migra para o oceano. Depois de dois anos na água salgada retornam para reproduzir e morrer. Neste período da fase adulta são pescados no oceano.

   - Para regulação da pesca, amostras de peixe são obtidas dos pescadores para identificação da origem (Alaska ou Canadá).

   - Os anéis de crescimento nas escamas do Salmão fornecem informação sobre o local de nascimento. Anéis associados ao crescimento em água doce são geralmente menores para o salmão do Alaska.

   - Informação coletada: origem (Alaska = 1 ou Canadá = 2), sexo (Fêmea = 1, Macho = 2), diâmetro de anéis para crescimento em água doce (Freshwater), diâmetro de anéis para cresciemento no oceano (Marine).

Exemplos

2. Johnson e Wichern, Exemplo 11.11 - Admissão de alunos em mestrado.

   - A admissão no Mestrado em Negócios utiliza um índice para auxiliar a decisão sobre admissão. Este índice combina a média de pontos na graduação (GPA) e desempenho no teste de admissão (GMAT).

   - Aplicações recentes foram classificadas em três grupos: aceitos, não aceitos e indecisão.

   - O objetivo é classificar um novo candidato (GPA = 3.21 e GMAT = 497) em um dos três grupos.

3. Iris Data - https://rpubs.com/Nolan/298913.

Discriminantes por Regressão Logística

GLMs para Dados Binários

Modelos para dados onde existem duas respostas possíveis - Sucesso (S) e Falha (F) - Distribuição de Bernoulli.

$$\mathrm{P(Y = 1) = P(S) = \pi} \quad e \quad \mathrm{P(Y = 0) = P(F) = 1 - \pi}$$

Para este modelo,

$$\mathrm{E(Y) = \pi} \quad e \quad \mathrm{Var(Y) = \pi(1 - \pi)}$$

A probabilidade de sucesso é escrita como $\pi(\mathrm{x})$ para indicar sua dependência de x.

Discriminantes por Regressão Logística

Modelo Probabilístico Linear

Um modelo simples relacionando π a x é modelo linear:

$$\pi(\mathrm{x}) = \alpha + \beta \mathrm{x}$$

Problemas com este modelo:

1. Para certos valores de x, $\pi(\mathrm{x}) > 1$ ou $\pi(\mathrm{x}) < 0$.
2. Solução de Mínimos Quadrados não é ótima porque $\mathrm{Var}(\mathrm{Y}) = \pi(\mathrm{x})(1 - \pi(\mathrm{x}))$.
3. Estimadores de Máxima Verossimilhança não tem forma fechada.

.

Discriminantes por Regressão Logística

Modelo de Regressão Logística

São modelos de regressão linear para dados binários. Tipicamente, estes modelos

- são monótonos com $\pi(\mathrm{x})$ - crescendo com x crescendo ou decrescendo com x decrescendo.
- satisfazem $0 \leq \pi(\mathrm{x}) \leq 1$.
- formam um curva em formato de S.

Um modelo satisfaz a função de regressão logística se

$$\text{logito}(\pi(\mathrm{x})) = \log\left(\frac{\pi(\mathrm{x})}{1-\pi(\mathrm{x})}\right) = \alpha + \beta \mathrm{x}$$

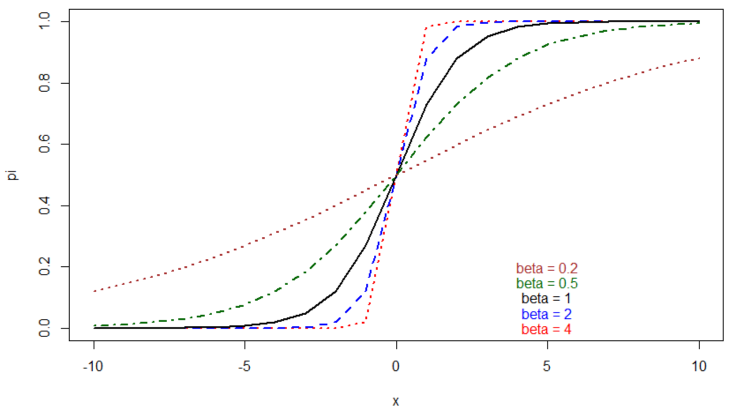
e podemos resolver para $\pi(x)$

$$\pi(\mathrm{x}) = \frac{\exp\{\alpha + \beta \mathrm{x}\}}{1 + \exp\{\alpha + \beta \mathrm{x}\}} = \frac{1}{1 + \exp\{-(\alpha + \beta \mathrm{x})\}}$$

Discriminantes por Regressão Logística

Exemplos de Curvas Logísticas

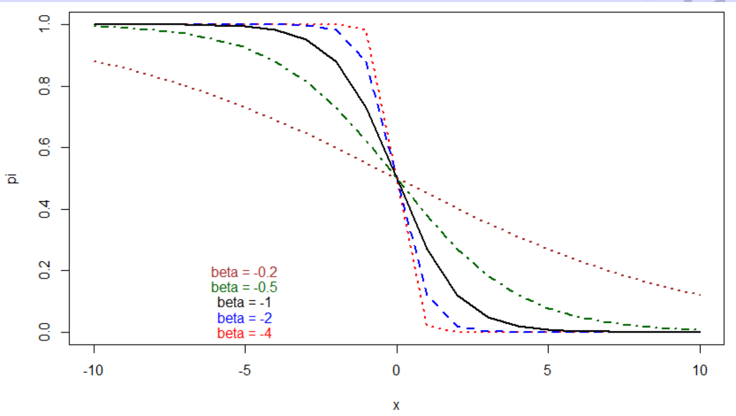
Curvas Logísticas com $\alpha = 0$ e β (positivo) variando.



Discriminantes por Regressão Logística

Exemplos de Curvas Logísticas

Curvas Logísticas com $\alpha = 0$ e β (negativo) variando.



- R possui a função `GLM` com a opção `FAMILY = BINOMIAL`.
- Detalhes: curso de dados categorizados.

Exemplos

4. Challenger Space Shuttle O-Ring Failures

- Data prevista de lançamento: 22 de Janeiro de 1986.
- Data do lançamento/acidente: 28 de Janeiro de 1986
  Cabo Canaveral, Flórida.
- Filme sobre o acidente:
  https://www.youtube.com/watch?v=2FehGJQlOf0
- Cobertura CBS news:
  https://www.youtube.com/watch?v=N9kWG-lAOCA
- Informações Gerais:
  http://pt.wikipedia.org/wiki/STS-51-L
- Livros:
  - McDonald, A.J. (2012) "Truth, Lies, and O-Rings: Inside the Space Shuttle Challenger Disaster". University Press of Florida.
  - Vaughan, D. (2016) "The Challenger Launch Decision: Risky Technology, Culture, and Deviance at NASA". University of Chicago Press.