Universidade Federal do Paraná

Laboratório de Estatística e Geoinformação - LEG

Classificação

Prof.: Eduardo Vargas Ferreira

Introdução

- Vimos até agora como prever o comportamento de uma variável **quantitativa** Y dado um vetor $\mathbf{x}$;
- Para tanto, estudamos métodos para encontrar $h(\mathbf{x})$ tal que o custo $J(h) = \mathrm{E}\left[(Y - h(\mathbf{x}))^2\right]$ fosse baixo;
- Em particular, vimos que $min\left\{J(y_i, h(\mathbf{x}))\right\} \approx min\left\{\frac{1}{n}\sum_{i=1}^{n}[y_i - h(x_i)]^2\right\}$;
- Por outro lado, em muitos problemas, a variável Y assume valores em um conjunto não ordenado $\mathcal{C}$, por exemplo:
  - $\star$ `E-mail` $\in \{$`spam`, `ham`$\}$;
  - $\star$ `Dígito` $\in \{0, 1, \ldots, 9\}$;
  - $\star$ `Alzheimer` $\in \{$`com Alzheimer`, `sem Alzheimer`$\}$;
- Nestes casos, estamos diante de um **problema de classificação**.

Introdução

- O que muda em um problema de classificação?
- O custo $J(h) = \mathrm{E}\left[(Y - h(\mathbf{x}))^2\right]$ não faz mais sentido;
- Ao invés dele, é comum utilizar

$$\begin{aligned} J(h) &= \mathrm{E}\left[\mathrm{I}(Y \neq h(\mathbf{X}))\right] \\ &= \mathrm{P}(Y \neq h(\mathbf{X})). \end{aligned}$$

- Por simplicidade, digamos que Y assuma somente dois valores, c_1 ou c_2 (um problema binário), então

$$\begin{aligned} J(h) &= \mathrm{P}(Y \neq h(\mathbf{X})) \\ &= \int_x \mathrm{P}(Y \neq h(\mathbf{X}|\mathbf{x}))f(\mathbf{x})d\mathbf{x} \\ &= \int_x \left[\mathrm{I}(h(\mathbf{x}) = c_2)\mathrm{P}(Y = c_1|\mathbf{x}) + \mathrm{I}(h(\mathbf{x}) = c_1)\mathrm{P}(Y = c_2|\mathbf{x})\right] f(\mathbf{x})d\mathbf{x} \end{aligned}$$

Classificador de Bayes

- Para um dado $\mathbf{x}$, devemos escolher $h(\mathbf{x}) = c_1$ quando

$$\mathrm{P}(Y = c_1|\mathbf{x}) \geq \mathrm{P}(Y = c_2|\mathbf{x}),$$

  caso contrário, $h(\mathbf{x}) = c_2$.

- Em outras palavras, a melhor h é dada por

$$h(\mathbf{x}) = \underset{d \in \{c_1, c_2\}}{\mathit{argmax}}\ \mathrm{P}(Y = d|\mathbf{x})$$

- Tal classificador é conhecido como **Classificador de Bayes**;

- Para o caso binário, ele é reescrito como

$$h(\mathbf{x}) = c_1 \iff \mathrm{P}(Y = c_1|\mathbf{x}) \geq \frac{1}{2}.$$

- No caso geral, a melhor h é dada por

$$h(\mathbf{x}) = \underset{c \in \mathcal{C}}{\mathit{argmax}}\ \mathrm{P}(Y = c|\mathbf{x})$$

Plug-in classifier

- O resultado anterior sugere uma abordagem simples para resolver o problema de predição

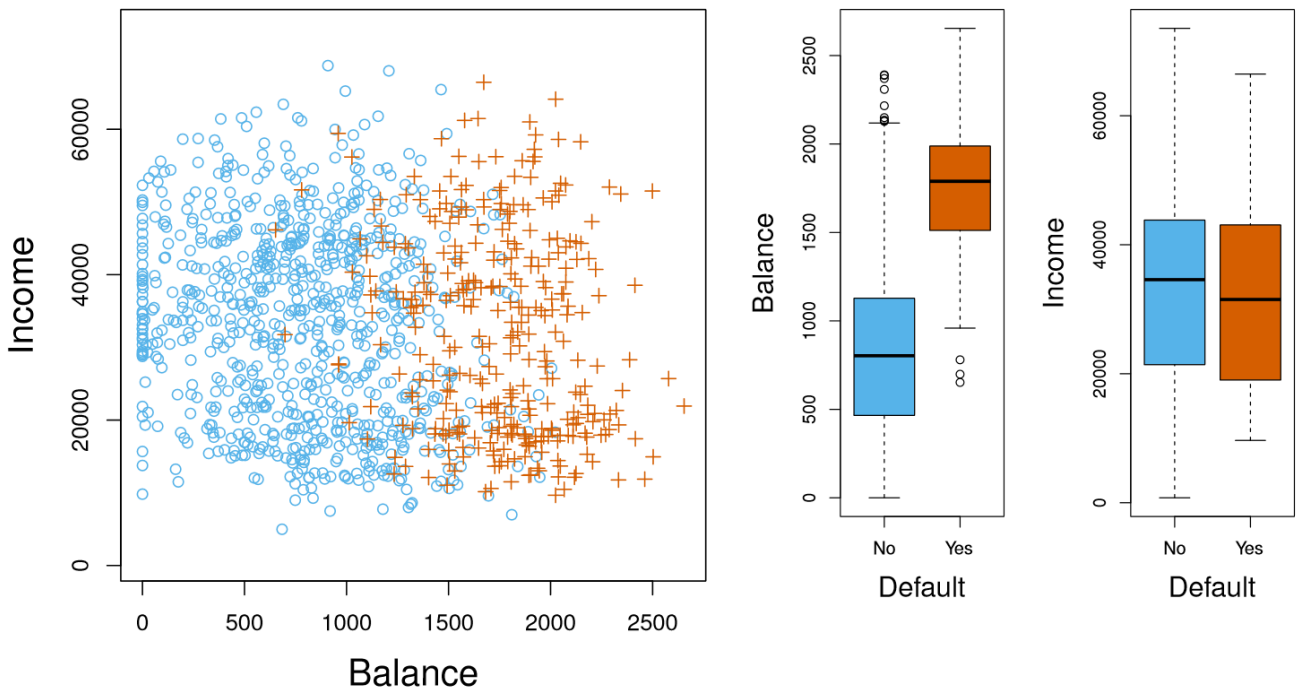
  ⋆ Estimamos $\mathrm{P}(Y = c|\mathbf{x})$ para cada categoria $c \in \mathcal{C}$;

  ⋆ Tomamos $h(\mathbf{x}) = \underset{c \in \mathcal{C}}{\mathit{argmax}}\ \widehat{\mathrm{P}}(Y = c|\mathbf{x})$.

- Esta abordagem é conhecida como **plug-in classifier**;
- Então, criar um classificador resume-se a estimar $\mathrm{P}(Y = c|\mathbf{x})$;
- Mas, como fazer isso? Veremos alguns métodos.

Exemplo: `Credit Card Default`

- Os dados referem-se aos rendimentos anuais do cartão de crédito (`income`) e saldos mensais (`balance`) de determinados clientes;
- Os indivíduos inadimplentes (`default`, *who defaulted*) em seus pagamentos são mostrados em laranja;



Podemos utilizar regressão linear?

- Suponha que para classificação da variável `default` codificamos da forma:

$$Y = \begin{cases} 0, & \text{se } \texttt{Não} \ , \\ 1, & \text{se } \texttt{Sim} \ . \end{cases}$$

- Podemos simplesmente realizar uma regressão linear de Y em X e classificar como `Sim` se $\hat{Y} > 0.5$?

  - ⋆ Considerando o fato de que $\mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = \mathrm{P}(Y = 1|\mathbf{X} = \mathbf{x})$, podemos pensar que regressão é ótima para isto!

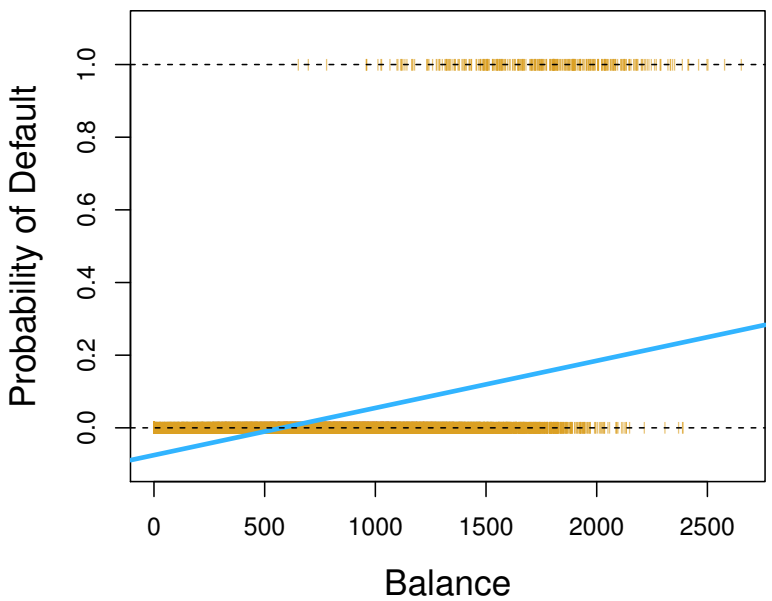
  - ⋆ No caso de resposta binária, regressão linear faz um bom trabalho (equivalente à **análise de discriminante linear**);

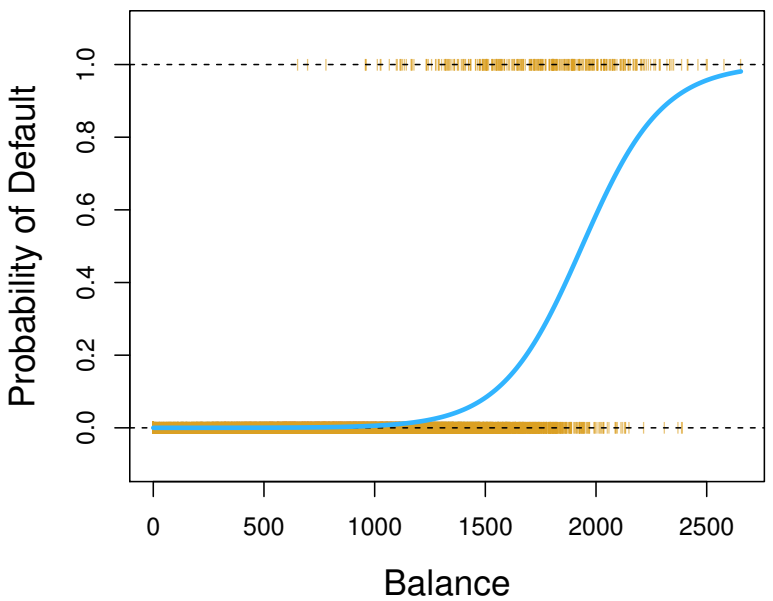
  - ⋆ Entretanto, ela pode produzir probabilidades menores do que 0 ou maiores do que 1.

  - ⋆ **Regressão logística** é mais apropriado.

Regressão linear versus logística

- As marcas em laranja indicam a resposta Y, 0 ou 1;
- No gráfico da esquerda a linha azul representa o ajuste através da regressão linear (algumas probabilidades estimadas são negativas!);
- As probabilidades preditas utilizando regressão logística (gráfico da direita) estão entre 0 e 1.





Regressão logística

- Por simplicidade, denotaremos $p(\mathbf{X}) = P(Y = 1|\mathbf{X})$. A regressão logística utiliza a forma

$$p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

- É fácil ver que, não importa os valores de β_0 e β_1 ou $\mathbf{X}$, $p(\mathbf{X})$ sempre assume valores entre 0 e 1;

- Note que, como no caso de regressão, não estamos assumindo que esta relação é verdadeira. Apenas que ela nos leva a um bom classificador;

- Com um pouco de algebrismo chegamos em

$$log\left(\frac{p(\mathbf{X})}{1 - p(\mathbf{X})}\right) = \beta_0 + \beta_1 X.$$

- Que é chamada **log odds** ou transformação **logit** em $p(\mathbf{X})$.

Regressão logística

- Utilizamos o método de máxima verossimilhança para estimar os parâmetros:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} [1 - p(x_i)]$$

- A verossimilhança fornece a probabilidade dos 0's e 1's observados nos dados. Encontramos os valores de β_0 e β_1 que maximiza esta função;

- Muitos pacotes estatísticos ajustam modelos de regressão logística por máxima verossimilhança. No `R` utilizamos a função `glm`

```
glm.fit = glm(formula,data = dados, family = binomial)
```

- Vamos voltar ao exemplo do `Credit Card Default`, considerando `balance` para prever `default`.

Credit Card Default: predições

Variável	Coeficiente	Erro padrão	Estatística t	p-valor
Intercepto	-10,6513	0,3612	-29,5	$< 0,0001$
Balance	0,0055	0,0002	24,9	$< 0,0001$

- Qual é a probabilidade estimada do default de alguém com balance de \$1000?

$$\hat{p}(\mathbf{X}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10,6513+0,0055\times 1000}}{1+e^{-10,6513+0,0055\times 1000}} = 0,006.$$

- E com balance de \$2000?

$$\hat{p}(\mathbf{X}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10,6513+0,0055\times 2000}}{1+e^{-10,6513+0,0055\times 2000}} = 0,586.$$

Credit Card Default: predições

- Vamos repetir o processo anterior, agora com `student` como preditor;

Variável	Coeficiente	Erro padrão	Estatística t	p-valor
`Intercepto`	-3,5041	0,0707	-49,55	$< 0,0001$
`Student[Yes]`	0,4049	0,1150	3,52	0,0004

$$\widehat{\mathrm{P}}\left(\texttt{default=Yes}|\texttt{student=Yes}\right) = \frac{e^{-3,5041+0,4049\times 1}}{1+e^{-3,5041+0,4049\times 1}} = 0,0431.$$

$$\widehat{\mathrm{P}}\left(\texttt{default=Yes}|\texttt{student=No}\right) = \frac{e^{-3,5041+0,4049\times 0}}{1+e^{-3,5041+0,4049\times 0}} = 0,0292.$$

Credit Card Default: várias variáveis

- Vamos considerar agora o caso de mais de um preditor. Assim, o modelo geral torna-se

$$log\left(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

e

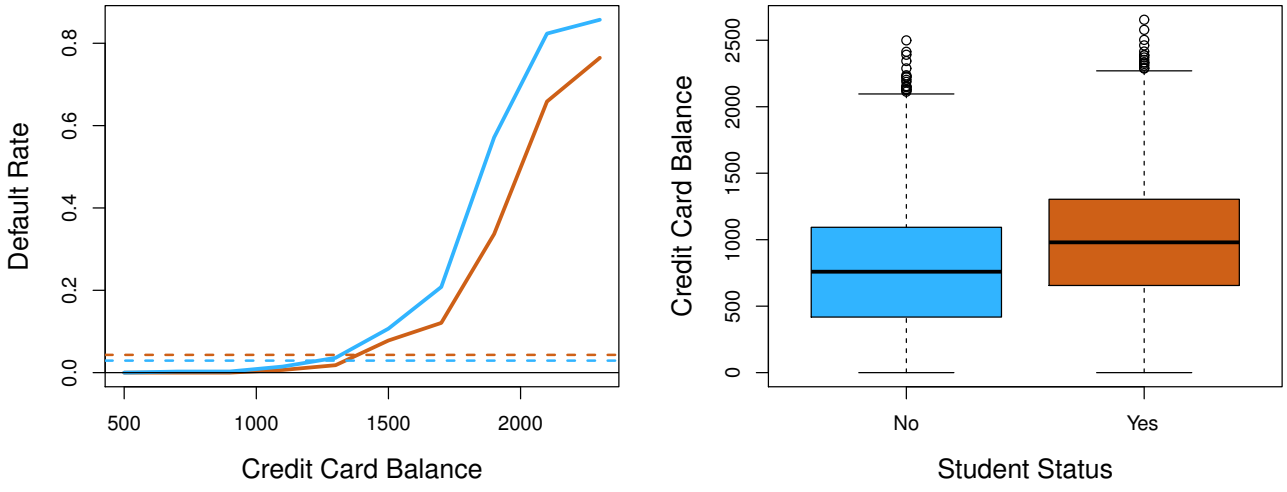
$$p(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\beta_p X_p}}.$$

Variável	Coeficiente	Erro padrão	Estatística t	p-valor
Intercepto	-10,8690	0,4923	-22,08	$< 0,0001$
Balance	0,0057	0,0002	24,74	$< 0,0001$
Income	0,0030	0,0082	0,37	0,7115
Student[Yes]	-0,6468	0,2362	-2,74	0,0062

- Por que o coeficiente de student é negativo agora, enquanto era positivo anteriormente? **Confundimento**.

Confundimento

- Os resultados são diferentes, especialmente quando existe correlação entre os preditores (veja o gráfico da direita);



- Estudantes tendem a ter maior saldo do cartão de crédito (`balance`);
- Assim, marginalmente a taxa de inadimplência (`default rate`) é maior do que não estudantes;
- Por outro lado, para cada nível do saldo mensal (`balance`) a inadimplência dos estudantes é menor (gráfico da esquerda).

Regressão multinomial

- Até agora, discutimos o caso de regressão logística com duas classes;
- É fácil generalizar para mais classes

$$P(Y = k|\mathbf{X}) = \frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{\sum_{l=1}^{K} e^{\beta_{0l}+\beta_{1l}X_1+\ldots+\beta_{pl}X_p}}$$

- Note que somente $K-1$ funções lineares são necessárias em uma regressão logística de K classes;
- O caso da regressão logística multiclasses é também chamada de regressão multinomial;
- Por exemplo, podemos classificar um paciente na sala de emergência de acordo com seu sintoma

$$Y = \begin{cases} 1, & \text{se } \texttt{AVC}\ , \\ 2, & \text{se } \texttt{overdose de droga}\ , \\ 3, & \text{se } \texttt{ataque epilético}\ . \end{cases}$$

Outra abordagem

- Uma alternativa para estimar $P(Y|X)$ consiste em modelar a distribuição de X em cada classe separadamente;
- E utilizar o **Teorema de Bayes** para obter $P(Y|X)$;

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{P(X = x)}$$

- Que escrevendo de outra forma fica

$$P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

- Então temos que

$$\delta_k(x) \propto \textit{argmax}\ \pi_k f_k(x)$$

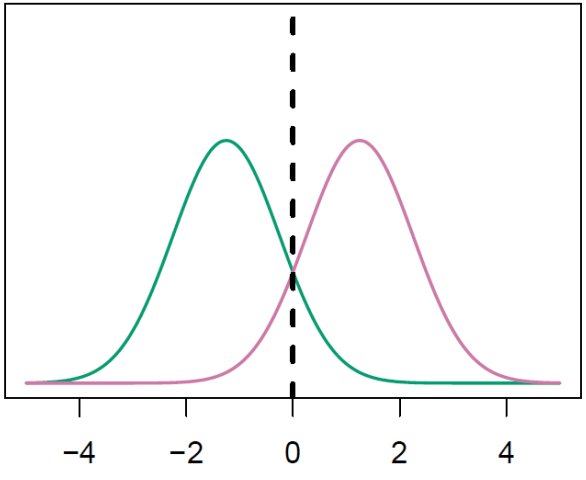
Outra abordagem

- $f_k(x) = P(\mathbf{X} = \mathbf{x}|Y = k)$ é a **densidade** para X na classe k (diferentes distribuições levam a diferentes métodos);

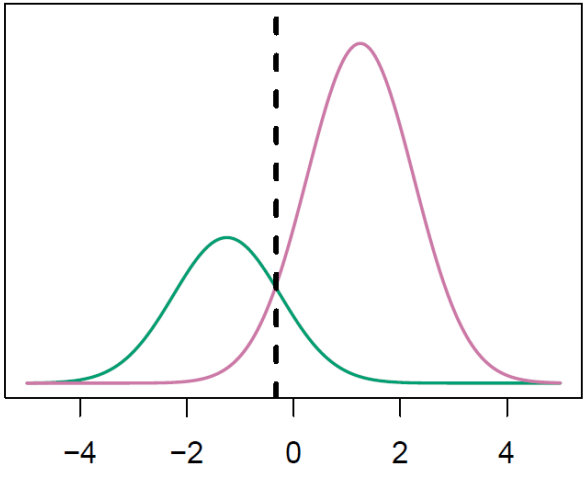
- $\pi_k = P(Y = k)$ é a **probabilidade marginal** ou **priori** para classe k. Pode ser estimada utilizando as proporções amostrais em cada classe.

- Com diferentes prioris em cada classe, temos diferentes tomadas de decisão;

π_1=.5, π_2=.5



π_1=.3, π_2=.7



- Por exemplo, o gráfico da direita a fronteira de decisão foi deslocada para esquerda.

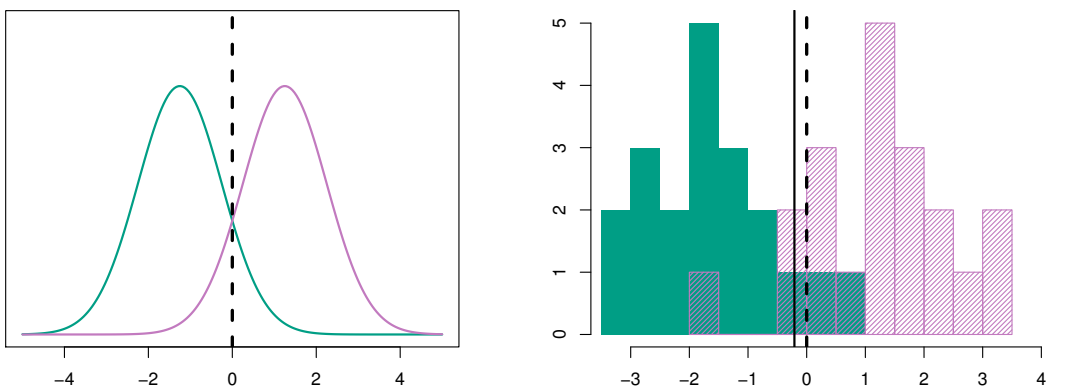
Análise de discriminante

- Ao considerarmos para $f_k(x)$ a distribuição Normal em cada classe, nos leva à **análise de discriminante linear** ou **quadrática**, pois

$$\begin{aligned}\delta_k(x) &\propto \textit{argmax}\ \pi_k f_k(x) \\ &= \textit{argmax}\left\{log\pi_k - \frac{1}{2}log|\Sigma_k| - \frac{1}{2}\langle x-\mu_k, \Sigma_k^{-1}(x-\mu_k)\rangle\right\}.\end{aligned}$$

- $\langle x-\mu_k, \Sigma_k^{-1}(x-\mu_k)\rangle$ é a **Distância de Mahalanobis** de x e μ_k;

- Por exemplo, seja $\mu_1 = -1.5, \mu_2 = 1.5, \pi_1 = \pi_2 = 0.5$ e $\sigma^2 = 1$

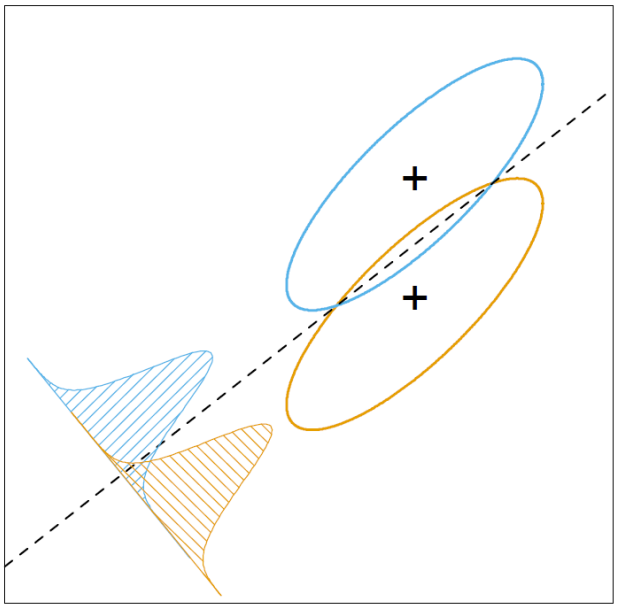


- Calculamos as distâncias entre x e μ_k (corrigido pelo $log\pi_k$), e classificamos de acordo com sua proximidade.

Análise de discriminante

- Ao contrário do exemplo anterior, não conhecemos os parâmetros das distribuições.

- Utilizamos, assim, os dados de treino para estimar tais quantidades e incorporar à regra de decisão, da seguinte forma



$$\hat{\pi}_k = \frac{n_k}{n}$$

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^t$$

Análise de discriminante

- Quando $f_k(x)$ possui matriz de covariância, $\mathbf{\Sigma}_k$, diferente em cada classe, temos a **análise de discriminante quadrático (ADQ)**

$$\begin{aligned}\delta_k(x) &\propto \textit{argmax}\ \pi_k f_k(x) \\ &= \textit{argmax}\left\{log\pi_k - \frac{1}{2}log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^t\Sigma_k^{-1}(x-\mu_k)\right\}.\end{aligned}$$

- Note a ocorrência do termo quadrático na distância de Mahalanobis;

- Se todas as classes compartilharem o mesmo $\Sigma = \sum_k \frac{n_k-1}{n-K}\hat{\Sigma}_k$, estamos diante da **análise de discriminante linear (ADL)**

$$\begin{aligned}\delta_k(x) &\propto \textit{argmax}\ \pi_k f_k(x) \\ &= \textit{argmax}\left\{log\pi_k - \frac{1}{2}\mu_k^t\Sigma^{-1}\mu_k + x^t\Sigma^{-1}\mu_k\right\}.\end{aligned}$$

- Agora, o termo quadrático foi cancelado.

Regressão logística versus ADL

- Regressão logística e análise de discriminante linear diferem-se na forma de estimar os parâmetros:
  - ⋆ Regressão logística maximiza a **verossimilhança condicional**

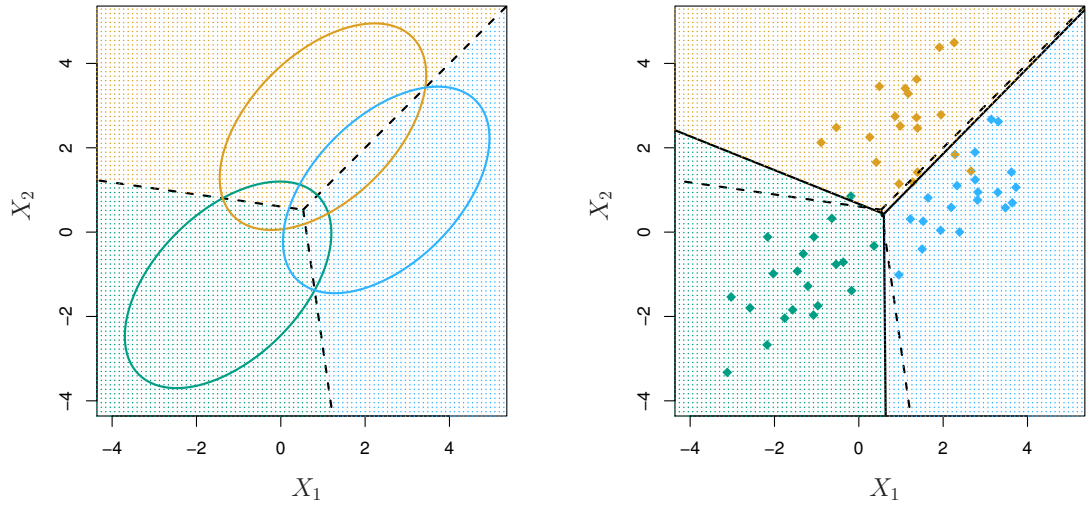
$$\prod_i p(x_i, y_i) = \underbrace{\prod_i p(y_i|x_i)}_{logistica} \underbrace{\prod_i g(x_i)}_{ignorado}$$

  - ⋆ ADL maximiza a **verossimilhança completa**

$$\prod_i p(x_i, y_i) = \underbrace{\prod_i p(x_i|y_i)}_{normal\ \ f_k} \underbrace{\prod_i p(y_i)}_{bernoulli\ \ \pi_k}$$

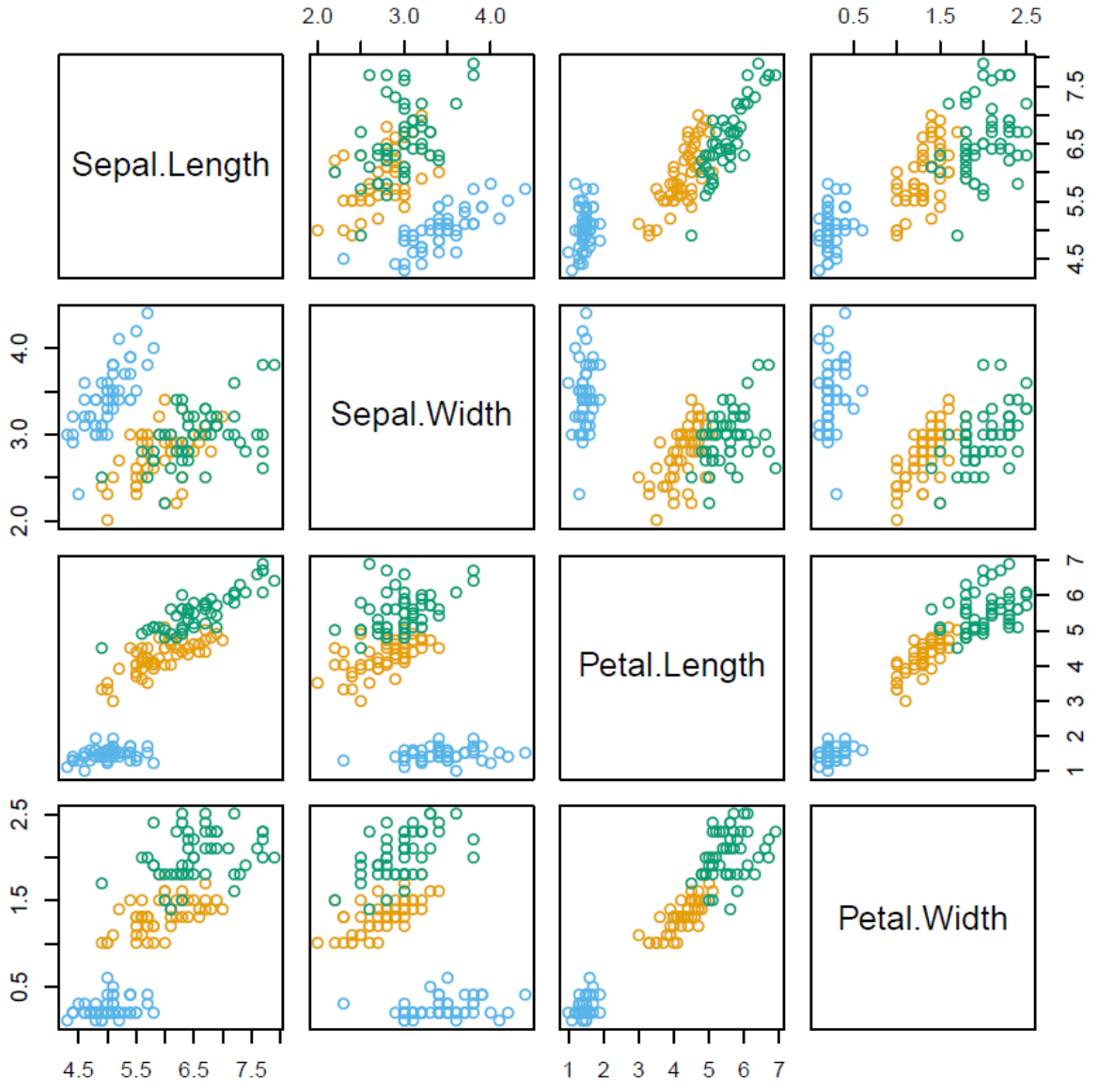
- Mas na prática, os resultados são similares.

Ilustração: $p = 2$ e $k = 3$ classes



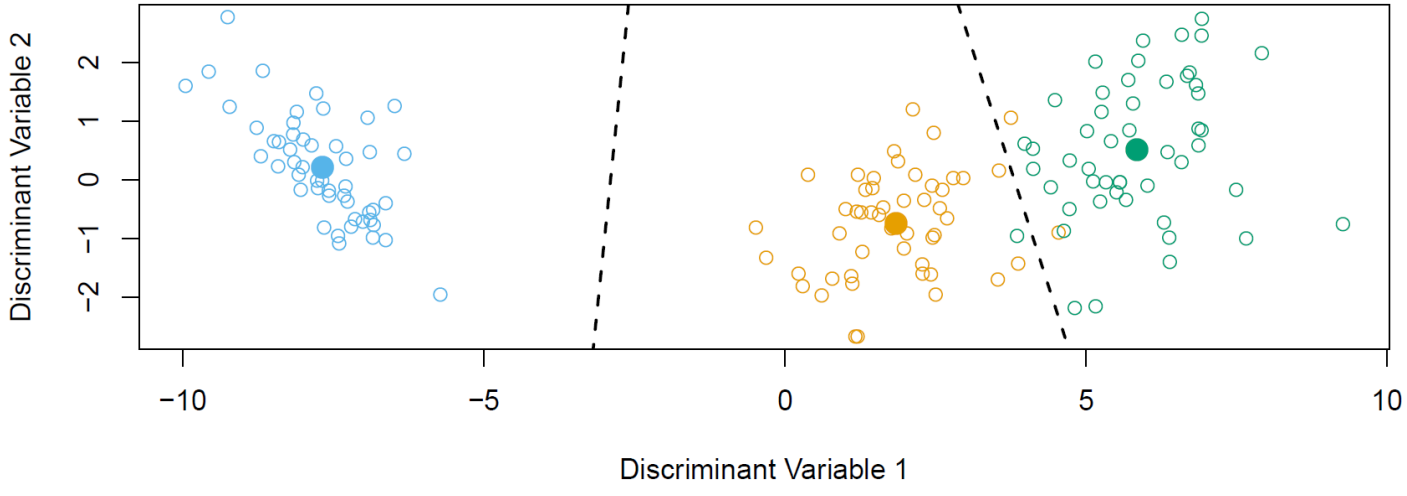
- Aqui, $\pi_1 = \pi_2 = \pi_3 = 1/3$;
- A linha pontilhada é conhecida como **fronteira de decisão de Bayes** (*Bayes decision boundaries*);
- Pode-se dizer que o objetivo de um classificador é se aproximar desta regra de decisão.

Exemplo: Iris Data



Exemplo: Iris Data

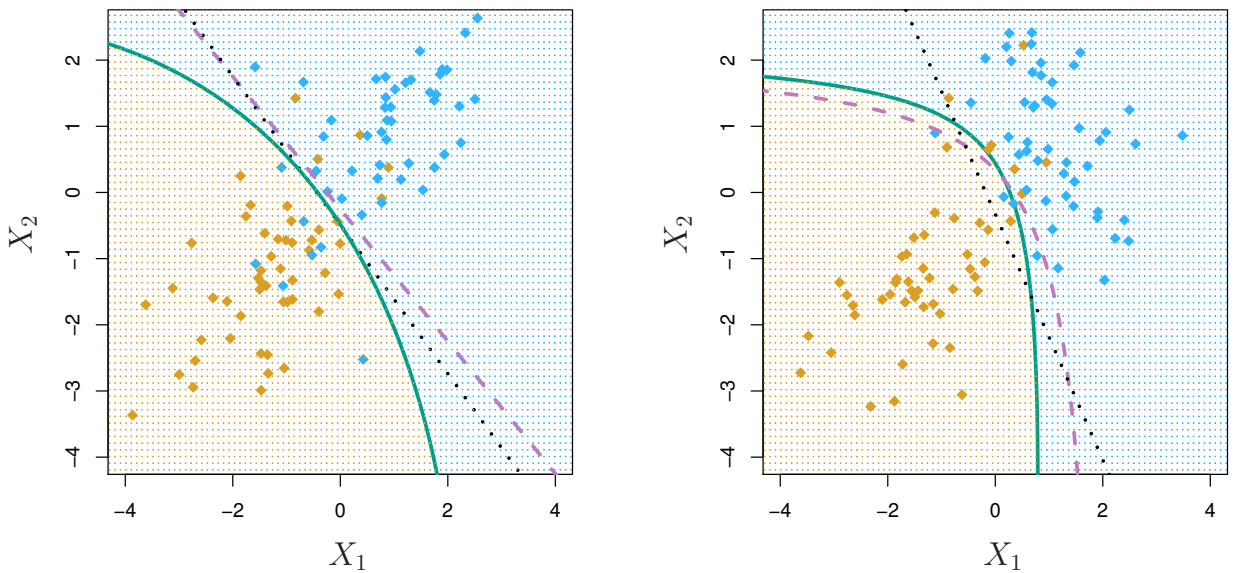
- Temos 4 variáveis, 3 espécies com 50 observações em cada classe;



- Análise de discriminante linear classifica corretamente 147/150 observações dos dados de treino.

Exemplo simulado: Bayes, ADL e ADQ

- No exemplo, temos a fronteira de decisão de Bayes em rosa, ADL pontilhado e ADQ em verde, em um problema com 2 classes;



- No gráfico da esquerda $\mathbf{\Sigma_1} = \mathbf{\Sigma_2}$;
- E no da direita $\mathbf{\Sigma_1} \neq \mathbf{\Sigma_2}$.

Tipos de erro: Credit Card Default

		Default verdadeiro		
		Não	Sim	Total
Default predito	Não	9644	252	9896
	Sim	23	81	104
	Total	9667	333	10000

Taxa de falso positivo: fração de exemplos negativos classificados como positivo - no exemplo temos 0,2% (23/9667);

Taxa de falso negativo: fração de exemplo positivo classificado como negativo - no exemplo temos 75,7% (252/333).

- Construímos esta tabela classificando a classe como Sim se

$$\widehat{\mathrm{P}}(\texttt{Default} = \texttt{Sim} | \texttt{Balance, Student}) \geq 0,5.$$

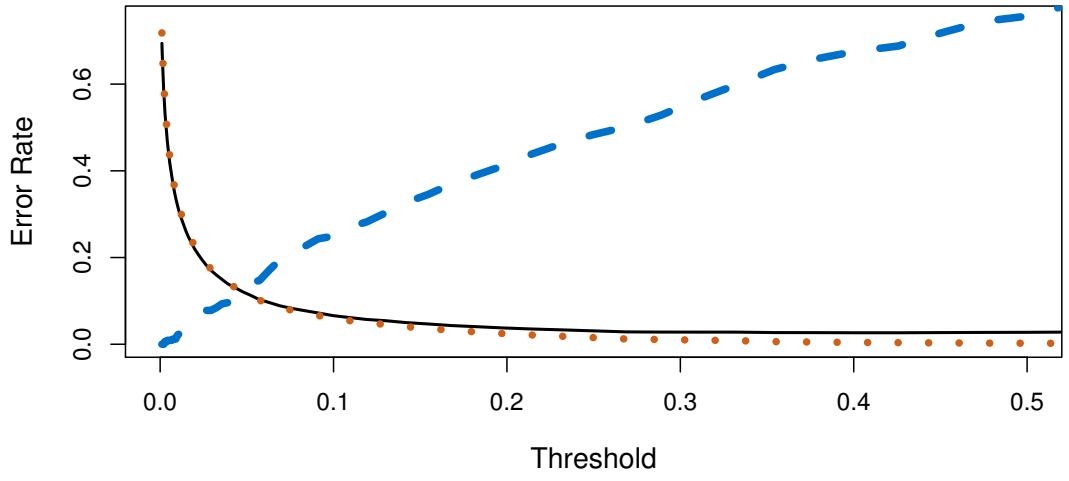
- Será que o limiar de 0,5 é a melhor opção?

Variando o threshold

- Podemos mudar as duas taxas de erro alterando a fronteira de decisão para algum valor $\in [0, 1]$:

$$\widehat{\mathrm{P}}(\texttt{Default} = \texttt{Sim} | \texttt{Balance, Student}) \geq \text{threshold}.$$

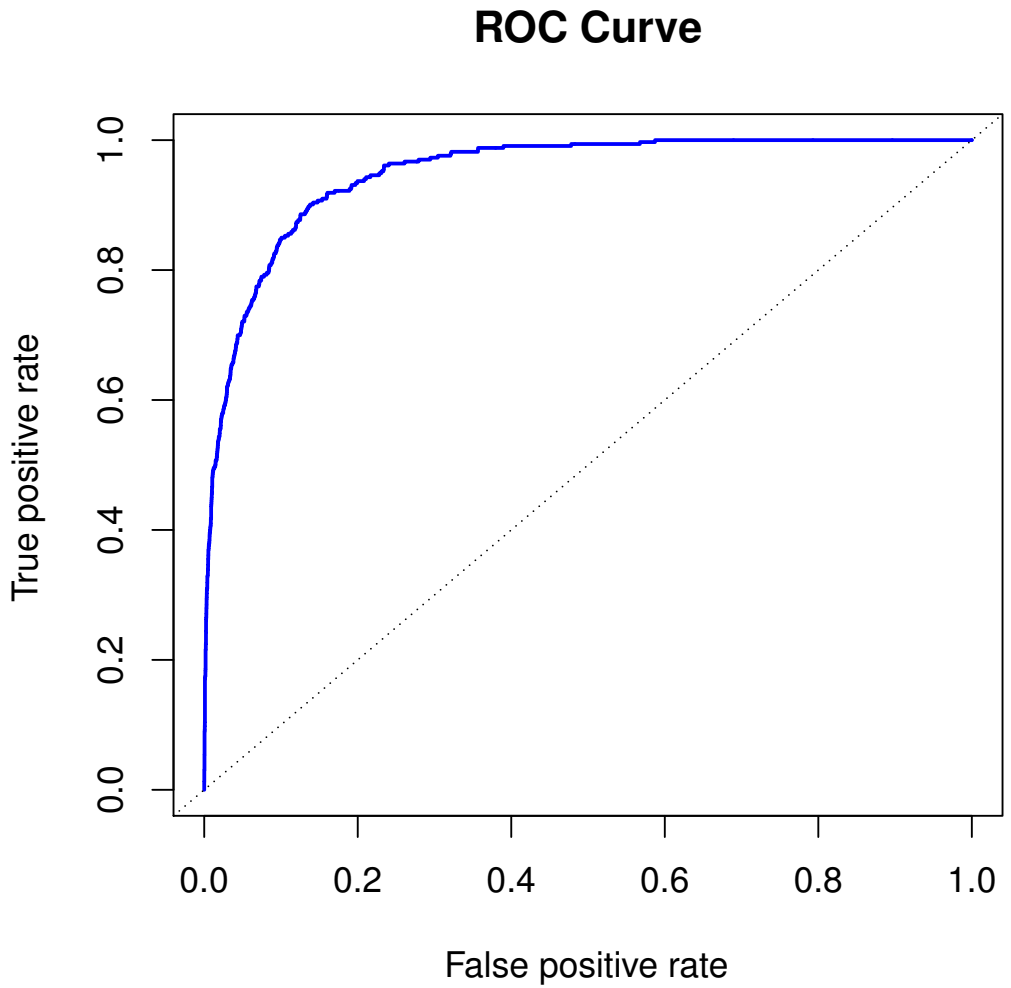
- A fim de reduzir a taxa de falso negativo, podemos reduzir o *threshold* para 0,1 (ou menos);



- Em azul temos a taxa de falso negativo, em laranja falso positivo e em preto a taxa de erro total.

Curva ROC

- A curva ROC (*receiver operator characteristic*) nos ajuda nesta escolha do *threshold*. Ela apresenta as duas taxas de erro ao mesmo tempo.



Naive bayes

- Vimos que quando $f_k(x)$ tem distribuição Normal com mesma variância $\mathbf{\Sigma}$ temos ADL. E se temos variâncias diferentes em cada classe temos ADQ;

- Agora, se supusermos que as componentes de $\mathbf{x}$ são independentes **condicionalmente à classe Y** estamos diante do **Naive Bayes**;

- Tal abordagem é útil quando p é grande e métodos multivariados como ADQ (mesmo ADL) enfrenta problemas;

- Naive Bayes normal assume $\mathbf{\Sigma}_k$ diagonal:

$$\delta_k(x) \propto log\left[\pi_k \prod_{j=1}^{p} f_{kj}(x_j)\right] = -\frac{1}{2}\sum_{j=1}^{p}\frac{(x_j-\mu_{kj})^2}{\sigma_{kj}^2} + log(\pi_k).$$

- Então, cada componente de $\mathbf{x}$ tem distribuição normal, com parâmetros que dependem da classe e da componente em questão;

- Apesar de tal suposição não ser razoável em muitos problemas (Naive = Ingênuo), ela é conveniente, e leva a bons classificadores.