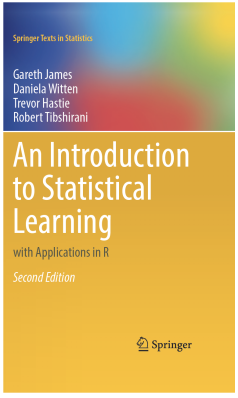
Aula 4: Métodos Lineares para Classificação

Machine Learning

Paulo Orenstein

Verão, 2025
IMPA



Capítulo 4: Métodos lineares para classificação

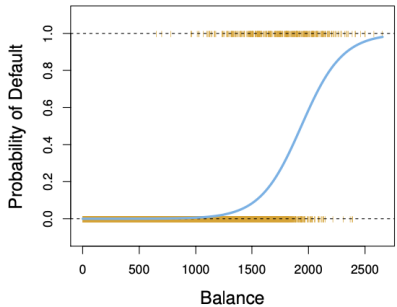
Créditos de figuras e slides: James, Witten, Hastie e Tibshirani

Capítulo 4

- Classificação: setup geral
- Métodos discriminativos: regressão logística e regressão multinomial
- Métodos generativos: LDA, QDA e naive Bayes
- Avaliação de classificadores
- Modelos lineares generalizados

Introdução

- Queremos prever variáveis qualitativas (*e.g.*, `cor de olho` $\in$ {`castanho`, `azul`, `verde`})
- Dado previsores X, nosso objetivo é construir $\hat{f}(X)$ tomando valores em $\mathcal{Y}$
- Em geral, há interesse não só na previsão, mas na probabilidade de pertencimento de cada classe



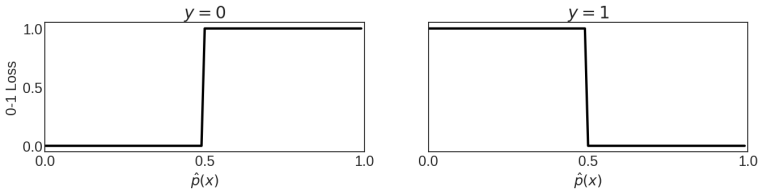
Setup geral

- Dados de treino: $\{(x_i, y_i)\}_{i=1}^n$, onde $x_i \in \mathcal{X}$ e $y_i \in \mathcal{Y}$
  - Vamos assumir que $\mathcal{Y} = \{0, 1\}$
- Classe de funções preditivas: $\hat{f}_{\text{tr}} : \mathcal{X} \to \mathcal{Y}$
  - Vamos assumir que $\hat{f}_{\text{tr}}(x) = \mathbb{I}_{[\hat{p}(x)>0.5]}$, com $\hat{p}(x) = h(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)$
- Função-perda: $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Vamos assumir que $L(y, \hat{f}(x)) = \mathbb{I}_{[y \neq \hat{f}(x)]}$
- Otimizador: encontrar $\hat{f}$ (ou, equivalentemente, $\hat{\beta}_0, \ldots, \hat{\beta}_p$) que minimiza o erro médio:

$$(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \underset{\tilde{\beta}_0, \ldots, \tilde{\beta}_p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[y_i \neq \hat{f}(x_i)]}$$

Classificação: otimização da perda 0-1

- Dado que temos uma função-perda, por que não escolher os parâmetros que minimizam a função?
- Problema: perda 0-1, $L(y, \hat{f}(x)) = \mathbb{I}_{[y \neq \hat{f}(x)]}$ com $\hat{f}(x) = \mathbb{I}_{[\hat{p}(x) > 0.5]}$, não é convexa, nem suave



- Ou seja, $\partial L(y, \hat{f}(x))/\partial \hat{p}(x) = 0$ em quase todo ponto!
- Daí, se $\hat{p}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$, temos $\partial L(y, \hat{f}(x))/\partial \hat{\beta}_j = (\partial L(y, \hat{f}(x))/\partial \hat{p}(x)) \cdot (\partial \hat{p}(x)/\partial \hat{\beta}_j)$
- Precisamos considerar perdas substitutas

Classificação: regressão linear

- Que tal usar a perda quadrática, como fizemos antes?
- Ou seja, por que não regressão linear com $Y \in \{0, 1\}$, e fitamos $\hat{p}(x) = \sum_{j=1}^{p} \hat{\beta}_j X_j$?
- Assim, $\mathbb{E}[Y|X = x] = 0 \cdot \mathbb{P}[Y = 0|X = x] + 1 \cdot \mathbb{P}[Y = 1|X = 1] = \mathbb{P}[Y = 1|X = x] = p(x)$
- Na hora de escolher a classe, basta tomar $\hat{f}(X) = \mathbb{I}_{[\hat{p}(x)>0.5]}$
- Ainda assim, há três desvantagens:
  - Não é imediato como generalizar para mais categorias (*e.g.*, {**avc**, **câncer**, **diabetes**})
  - As estimativas $\hat{p}(x) = \sum_{i=1}^{p} \hat{\beta}_j X_j$ não precisam estar em $[0, 1]$, então não são probabilidades
  - Se $y = 1$ e $\hat{p}(x) = 3$, o que é "bom", a perda seria menor com $\hat{p}(x) = 0$, pois $(3-1)^2 > (0-1)^2$
- Parte do problema é que estamos adaptando um método de regressão que assume $y \in \mathbb{R}$
- Dados sugerem repensar nossa função perda L e, com isso, nossa função preditiva $\hat{f}$

Regressão logística

- Vamos modelar $Y \sim \text{Bern}(p(x))$, ou seja, $\mathbb{P}[Y = 1|X = x] = p(x)$
- Essa formulação sugere uma perda natural: máxima verossimilhança, ou seja, escolher o valor de $\hat{p}(x)$ tal que os dados se pareçam com $Y \sim \text{Bern}(\hat{p}(x))$. Para isso note que

$$\mathbb{P}[Y = y|X = x] = (p(x))^y(1 - p(x))^{1-y}$$

- Isso vira uma perda se considerarmos o seu negativo (e tomarmos o log, por simplicidade):

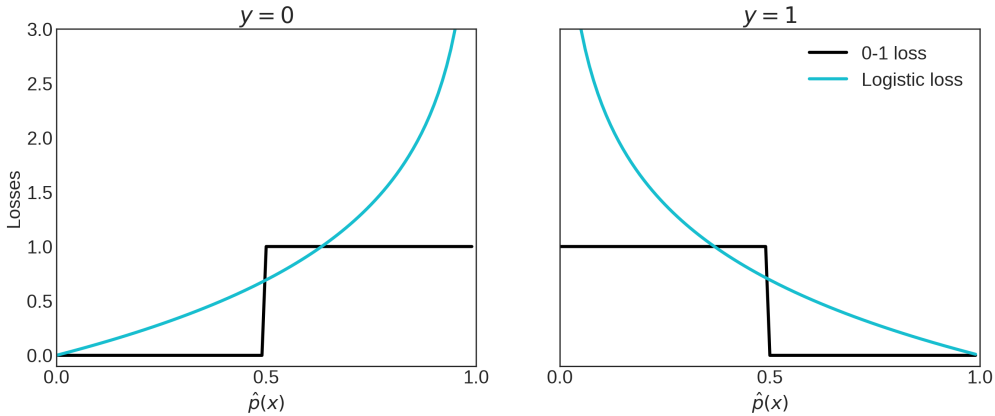
$$L(y, \hat{p}(x)) = -\left(y \log(\hat{p}(x)) + (1 - y) \log(1 - \hat{p}(x))\right)$$

- Essa perda é um substituto mais suave para a acurácia: $L(y, \hat{f}(x)) = \mathbb{I}_{[y \neq \hat{f}(x)]} = \mathbb{I}_{[y \neq \mathbb{I}_{[\hat{p}(x) > 0.5]}]}$
- Melhor: ele penaliza bastante as previsões feitas com muita confiança, mas erradas

Capítulo 4

- Classificação: setup geral
- Métodos discriminativos: regressão logística e regressão multinomial
- Métodos generativos: LDA, QDA e naive Bayes
- Avaliação de classificadores
- Modelos lineares generalizados

Regressão logística



Regressão logística

- Se queremos escolher $\hat{f}(x)$ para minimizar a perda empírica, então:

$$\hat{p}(x) = \underset{\tilde{p}}{\operatorname{argmin}} - \left(\frac{1}{n} \sum_{i=1}^{n} y_i \log \frac{\tilde{p}(x_i)}{1 - \tilde{p}(x_i)} + \log(1 - \tilde{p}(x_i)) \right)$$

- Ideia: vamos modelar o impacto sobre y_i de maneira linear, ou seja:

$$\log \frac{\hat{p}(x)}{1 - \hat{p}(x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- Isso é equivalente a tomar:

$$\hat{p}(x) = \text{logistic}(\beta_0 + \cdots + \beta_p x_p) = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p}},$$

  o que resolve nosso problema com previsões fora de $[0, 1]$. É uma normalização.

Regressão logística

- Ou seja: acurácia é uma perda desejada, mas difícil de otimizar. A logística é mais suave.
- Escolhemos $\hat{p}(x_i)$ para ter impacto linear dos previsores sobre y_i. É uma ideia bem geral.
- Isso equivale a resolver o seguinte problema de estimação:

$$
\begin{aligned}
(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p) &= \underset{\tilde{\beta}_0,\ldots,\tilde{\beta}_p}{\operatorname{argmin}} -\frac{1}{n}\sum_{i=1}^{n} y_i \log \frac{\hat{p}(x_i)}{1-\hat{p}(x_i)} + \log(1-\hat{p}(x_i)) \\
&= \underset{\tilde{\beta}_0,\ldots,\tilde{\beta}_p}{\operatorname{argmin}} -\frac{1}{n}\sum_{i=1}^{n} y_i \beta^T x_i - \log(1+e^{\beta^T x_i})
\end{aligned}
$$

- É um problema convexo. Podemos usar métodos de gradiente para encontrar β.

Regressão logística: otimização

- Em mais detalhes, note que função-perda é convexa em β:

$$L = -\frac{1}{n}\sum_{i=1}^{n} y_i \beta^T x_i - \log(1 + e^{\beta^T x_i})$$

- Logo, podemos usar métodos de gradiente baseados em:

$$\frac{\partial L}{\partial \beta} = -\left(\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{x_i e^{\beta^T x_i}}{1 + e^{\beta^T x_i}}\right) = -\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{p}(x_i))x_i = -\frac{1}{n}X^T(y - \hat{p})$$

$$\frac{\partial^2 L}{\partial \beta \partial \beta^T} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T(\hat{p}(x_i) - \hat{p}^2(x_i)) = \frac{1}{n}X^T W X,$$

onde $W = \text{diag}(\hat{p}(1 - \hat{p}))$

Regressão logística: otimização

- Descida de gradiente com passo η (método de primeira ordem):

$$\hat{\beta}_{t+1} = \hat{\beta}_t - \eta_t \frac{\partial L}{\partial \beta} = \hat{\beta}_t + \frac{\eta_t}{n} X^T (y - \hat{p})$$

- Método de Newton (método de segunda ordem):

$$\begin{aligned}\hat{\beta}_{t+1} &= \hat{\beta}_t - \left(\frac{\partial^2 L}{\partial \beta \partial \beta^T} \right)^{-1} \frac{\partial L}{\partial \beta} = \hat{\beta}_t + (X^T W X)^{-1} X^T (y - \hat{p}) \\ &= (X^T W X)^{-1} X^T W (X\hat{\beta}_t - W^{-1}(y - \hat{p})) = (X^T W X)^{-1} X^T W z_t,\end{aligned}$$

  onde $z_t = X\hat{\beta}_t + W^{-1}(y - \hat{p})$

- Isso é uma forma de mínimos quadrados iterados com pesos (IRLS), onde a i-ésima observação recebe peso $\sqrt{\hat{p}_i(1 - \hat{p}_i)}$; aí cada passo de Newton pode ser resolvido com descida de gradiente:

$$\hat{\beta}_{t+1} = \underset{\tilde{\beta}}{\operatorname{argmin}} (z_t - X\tilde{\beta})^T W (z_t - X\tilde{\beta})$$

Regressão logística: exemplo

- Resultado da regressão logística **default** $\approx$ Bern(logistic($\beta_0 + \beta_1$**balance**)):

	Coefficient	Std. Error	Z-statistic	p-value
Intercept	-10.6513	0.3612	-29.5	<0.0001
balance	0.0055	0.0002	24.9	< 0.0001

- Inferência pode ser generalizada através de testes assintóticos

- Qual é a probabilidade de **default** para alguém com **balance** de \$1000?

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

- Qual é a probabilidade de **default** para alguém com **balance** de \$2000?

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Regressão logística: exemplo

- Resultado da regressão logística **default** $\approx$ Bern(logistic($\beta_0 + \beta_1$ **student**)):

	Coefficient	Std. Error	Z-statistic	p-value
Intercept	-3.5041	0.0707	-49.55	<0.0001
student[Yes]	0.4049	0.1150	3.52	0.0004

- Qual é a probabilidade de **default** para um estudante?

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

- Qual é a probabilidade de **default** para um não-estudante?

$$\hat{p}(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x}} = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Regressão logística com muitas variáveis

- O caso com p previsores é estimado como antes, via máxima verossimilhança:

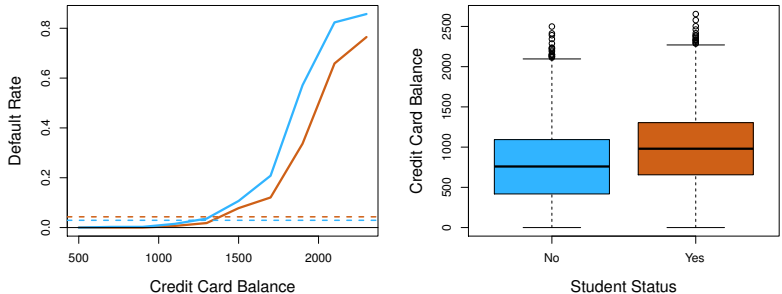
$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p}}$$

- Resultado de $\texttt{default} \approx \text{Bern}(\text{logistic}(\beta_0 + \beta_1 \texttt{balance} + \beta_2 \texttt{income} + \beta_3 \texttt{student}))$:

	Coefficient	Std. Error	Z-statistic	p-value
`Intercept`	-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	0.2363	-2.74	0.0062

- Por que o coeficiente `student`, que antes era positivo, agora ficou negativo?

Regressão logística: variáveis de confusão



- Estudantes têm saldos maiores que não-estudantes, então saldo é explicado por estudante
- Pessoas com grandes saldos têm maior probabilidade de inadimplência
- Dentre pessoas com alto saldo, estudantes são menos prováveis de serem inadimplentes

Regressão logística: problemas

- Quando há colinearidade, os coeficientes ficam instáveis
- Quando há separação perfeita entre classes, os coeficientes ficam instáveis
  - É sempre o caso com $p \geq n - 1$
- Coeficientes instáveis tornam o problema de otimização mais complicado
- A interpretação dos coeficientes não é tão simples quanto no caso de regressão linear
- O modelo pode ser muito inflexível para alguns problemas, como classificação de dígitos

Regressão multinomial

- Como proceder se queremos prever mais de duas classes?
- Por exemplo, **emergências** $\in \{$**infarto**, **overdose**, **epilepsia**$\}$
- Ideia: utilizar a distribuição categórica ao invés da Bernoulli. Com K classes, suponha que $\mathbb{P}[Y = k|X = x] = \hat{p}_k(x)$, onde $\sum_{k=1}^{K} \hat{p}_k(x) = 1$. Ou seja, a densidade de Y é dada por

$$\mathbb{P}[Y = y|X = x] = \hat{p}_1(x)^{\mathbb{I}_{[y=1]}} \cdot \hat{p}_2(x)^{\mathbb{I}_{[y=2]}} \cdots \hat{p}_K(x)^{\mathbb{I}_{[y=K]}}$$

- A log-verossimilhança é, portanto,

$$\begin{aligned} L(y, \hat{p}(x)) &= \sum_{k=1}^{K} \mathbb{I}_{[y=k]} \log \hat{p}_k(x) = \sum_{k=1}^{K-1} \mathbb{I}_{[y=k]} \log \frac{\hat{p}_k(x)}{\hat{p}_K(x)} + \log \hat{p}_K(x) \\ &= \sum_{k=1}^{K-1} \mathbb{I}_{[y=k]} \log \left(\frac{\hat{p}_k(x)}{1 - \sum_{l=1}^{K-1} \hat{p}_l(x)} \right) + \log \left(1 - \sum_{k=1}^{K-1} \hat{p}_k(x) \right) \end{aligned}$$

Regressão multinomial

- Como antes, modelamos o efeito de $\hat{p}(x)$ em y de maneira linear, ou seja,

$$\log\left(\frac{\hat{p}_k(x)}{1-\sum_{k=1}^{K-1}\hat{p}_l(x)}\right) = \hat{\beta}_{k0} + \hat{\beta}_{k1}x_1 + \hat{\beta}_{k2}x_2 + \cdots + \hat{\beta}_{kp}x_p,$$

onde agora há $(p+1)K$ parâmetros. Ou seja, tomamos

$$\hat{p}_k(x) = \frac{e^{\hat{\beta}_{k0}+\hat{\beta}_{k1}x_1+\hat{\beta}_{k2}x_2+\cdots+\hat{\beta}_{kp}x_p}}{1+\sum_{l=1}^{K-1} e^{\hat{\beta}_{l0}+\hat{\beta}_{l1}x_1+\hat{\beta}_{l2}x_2+\cdots+\hat{\beta}_{lp}x_p}}, \qquad k = 1, \ldots, K-1$$

e, para a última classe,

$$\hat{p}_K(x) = 1 - \sum_{k=1}^{K-1}\hat{p}_k(x) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\hat{\beta}_{l0}+\hat{\beta}_{l1}x_1+\hat{\beta}_{l2}x_2+\cdots+\hat{\beta}_{lp}x_p}}$$

- Os parâmetros são determinados via máxima verossimilhança, como antes

Capítulo 4

- Classificação: setup geral
- Métodos discriminativos: regressão logística e regressão multinomial
- Métodos generativos: LDA, QDA e naive Bayes
- Avaliação de classificadores
- Modelos lineares generalizados

Modelos generativos

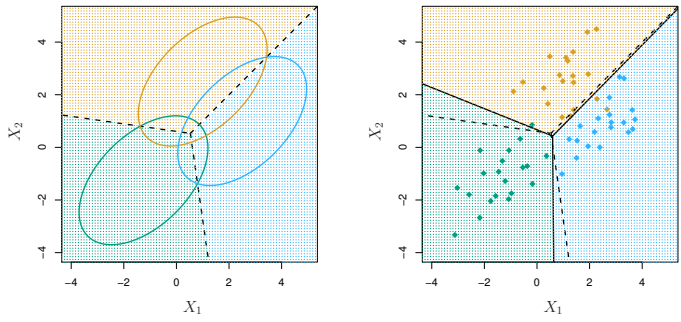
- Até agora, o que fizemos foi estimar $\mathbb{P}[Y = y|X = x] = p(x)$.
- Agora, vamos fazer isso estimando outras duas quantidades:
  1. $\mathbb{P}[X|Y]$: dada a categoria de Y, qual é a distribuição sobre os previsores X
  2. $\mathbb{P}[Y]$: a probabilidade de Y pertencer a cada uma das categorias
- Daí usamos a regra de Bayes, que diz:

$$\mathbb{P}[Y = k|X = x] = \frac{\mathbb{P}[X = x|Y = k]\ \mathbb{P}[Y = k]}{\mathbb{P}[X = x]} = \frac{\mathbb{P}[X = x|Y = k]\ \mathbb{P}[Y = k]}{\sum_{l=1}^{K} \mathbb{P}[X = x|Y = l]\ \mathbb{P}[Y = l]}$$

- Como vamos estimar $\mathbb{P}[X|Y]$ e $\mathbb{P}[Y]$, esses modelos são chamados generativos
  - Modelos que estimam $\mathbb{P}[Y|X]$ (como regressão logística) são chamados discriminativos

Análise discriminante linear

- Vamos estimar $\mathbb{P}[Y|X]$ usando as seguintes hipóteses:
  1. $\mathbb{P}[X = x|Y = k] = \hat{f}_k(x)$ é uma distribuição Normal multivariada



  2. $\mathbb{P}[Y = k] = \hat{\pi}_k$ é exatamente a fração de dados de treino na classe k

Análise discriminante linear

- Seja $\mathbb{P}[Y = k] = \pi_k$
- Suponha que $\mathbb{P}[X = x|Y = k]$ é uma Normal multivariada:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k)},$$

  onde $\mu_k \in \mathbb{R}^p$ é a média da categoria k e $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ é a matriz de covariância (comum a todas as categorias)

- O classificador de Bayes para escolher a classe k é dado por:

$$\mathbb{P}[Y = k|X = x] = \frac{\mathbb{P}[X = x|Y = k]\mathbb{P}[Y = k]}{\mathbb{P}[X = x]} = \frac{f_k(x)\pi_k}{\mathbb{P}[X = x]} = C \times f_k(x)\pi_k$$

Análise discriminante linear

- Expandindo $f_k(x)$ e absorvendo o que não depende de k:

$$\mathbb{P}[Y = k|X = x] = \frac{C\pi_k}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k)} = \tilde{C}\pi_k e^{-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k)}$$

- Daí, dado X, vamos escolher a classe $Y = k$ que maximiza

$$\begin{aligned}\log \mathbb{P}[Y = k|X = x] &= \log \tilde{C} + \log \pi_k - \frac{1}{2}(x - \mu_k)^T\mathbf{\Sigma}^{-1}(x - \mu_k) \\ &= \log \tilde{\tilde{C}} + \log \pi_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + x^T\mathbf{\Sigma}^{-1}\mu_k\end{aligned}$$

  e note que o termo $\log \tilde{\tilde{C}}$ não importa para k

- Dado x, nossa previsão é a classe k com maior valor

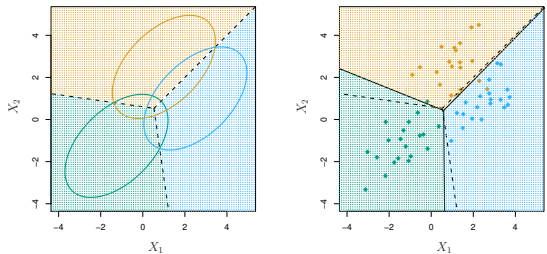
$$\delta_k(x) = \log \pi_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + x^T\mathbf{\Sigma}^{-1}\mu_k$$

Análise discriminante linear

- O "linear" se refere ao formato da fronteira de classificação
- O conjunto de pontos x igualmente classificados para as classes k e l são:

$$\log \pi_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + x^T \mathbf{\Sigma}^{-1}\mu_k = \delta_k(x) = \delta_l(x) = \log \pi_l - \frac{1}{2}\mu_l^T \mathbf{\Sigma}^{-1}\mu_l + x^T \mathbf{\Sigma}^{-1}\mu_l$$

- Isso é linear em x:



Análise discriminante linear: estimação

- Quais parâmetros precisamos estimar para LDA?
- $\hat{\pi}_k$ é a fração de amostras de treino na classe k:

$$\hat{\pi}_k = \frac{\#\{i : y_i = k\}}{n}$$

- $\hat{\mu}_k$ é o centro de massa de cada classe k:

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

- $\hat{\mathbf{\Sigma}}$ é o estimador não-viesado da matriz de covariância:

$$\hat{\mathbf{\Sigma}} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Análise discriminante linear: estimação

- Ou seja, tendo estimado $\hat{\pi}_k$, $\hat{\mu}_k$ e $\hat{\mathbf{\Sigma}}$, classificamos a resposta do previsor x como

$$\hat{f}(x) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \hat{\delta}_k(x) = \underset{k=1,\ldots,K}{\operatorname{argmax}} \log \hat{\pi}_k - \frac{1}{2}\hat{\mu}_k^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k + x^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k$$

- Por que usar análise discriminante linear?
  - Quando as classes são bem-separadas, LDA é mais estável de se estimar que regressão logística
  - Com n pequeno e $X|Y$ distribuído de maneira aproximadamente Normal, LDA é melhor que regressão logística

Análise discriminante: generalizações

- A hipótese fundamental de LDA é estimar

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

  onde $f_k(x)$ são Normais com mesma matriz de covariância $\mathbf{\Sigma}$ em cada classe.

- Considerando outros formatos para $f_k(x)$ é possível generalizar LDA:

  - Análise discriminante quadrática (QDA): quando cada classe tem um $\mathbf{\Sigma}_k$ diferente, ou seja, $X|Y = k \sim N(\mu_k, \mathbf{\Sigma}_k)$. Isso generaliza e é mais flexível do que LDA

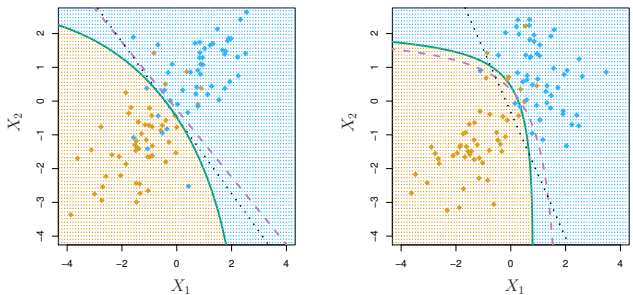
  - Naive Bayes: quando os p previsores são considerados independentes, *i.e.*, $f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)$ (e cada f_{kj} precisa ser especificado)

Análise discriminante quadrática

- Como $\mathbf{\Sigma}_k$ são diferentes, os termos quadráticos importam:

$$\delta_k(x) = \log \pi_k - \frac{1}{2}(x - \mu_k)^T \mathbf{\Sigma}_k^{-1}(x - \mu_k) - \frac{1}{2} \log |\mathbf{\Sigma}_k|$$

- Exemplo: fronteira de Bayes (– – –), LDA (· · · · · ·) e QDA (——)



Naive Bayes

- Assumindo que os p previsores são independentes é uma forma de simplificar o problema, útil quando p é grande
- Exemplo: no caso Normal, $\mathbf{\Sigma}_k$ é diagonal, então

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

- Naive Bayes pode ser usado também para variáveis quantitativas: basta usar uma estimativa da função de probabilidade (*e.g.*, histogramas)
- A hipótese de independência condicional de Naive Bayes pode ser forte, mas costuma funcionar

Comparando os métodos: analiticamente

- Dado previsor X, escolhemos a classe que maximiza $\mathbb{P}[Y = k|X = x]$ ou $\log\left(\frac{\mathbb{P}[Y=k|X=x]}{\mathbb{P}[Y=K|X=x]}\right)$
- Regressão logística:

$$\log\left(\frac{\mathbb{P}[Y = k|X = x]}{\mathbb{P}[Y = K|X = x]}\right) = \beta_{k0} + \sum_{j=1}^{p} \beta_{kj} x_j$$

- Análise discriminante linear:

$$\log\left(\frac{\mathbb{P}[Y = k|X = x]}{\mathbb{P}[Y = K|X = x]}\right) = \log\left(\frac{\pi_k \exp(-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k))}{\pi_K \exp(-\frac{1}{2}(x-\mu_K)^T \mathbf{\Sigma}^{-1}(x-\mu_K))}\right) = a_k + \sum_{j=1}^{p} b_{kj} x_j$$

Comparando os métodos: analiticamente

- Análise discriminante quadrática:

$$\log\left(\frac{\mathbb{P}[Y=k|X=x]}{\mathbb{P}[Y=K|X=x]}\right) = \log\left(\frac{\pi_k \exp(-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma_k}^{-1}(x-\mu_k))}{\pi_K \exp(-\frac{1}{2}(x-\mu_K)^T \mathbf{\Sigma_k}^{-1}(x-\mu_K))}\right)$$
$$= a_k + \sum_{j=1}^{p} b_{kj} x_j + \sum_{j=1}^{p}\sum_{l=1}^{p} c_{kjl} x_j x_l$$

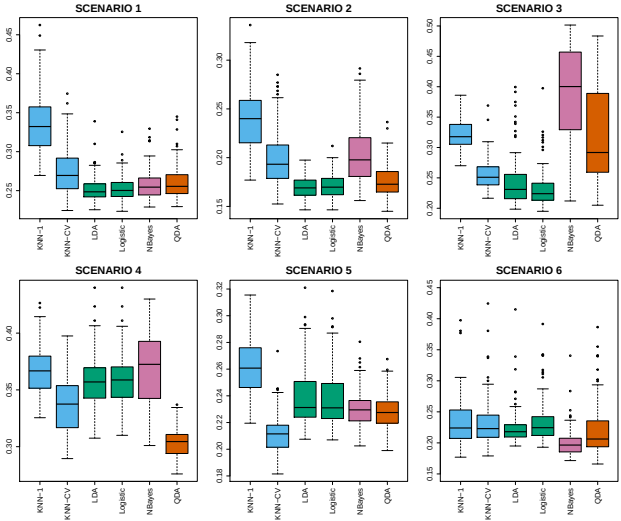
- Naive Bayes:

$$\log\left(\frac{\mathbb{P}[Y=k|X=x]}{\mathbb{P}[Y=K|X=x]}\right) = \log\left(\frac{\pi_k \prod_{j=1}^{p} f_{kj}(x)}{\pi_K \prod_{j=1}^{p} f_{Kj}(x_j)}\right) = a_k + \sum_{k=1}^{p} g_{kj}(x_j)$$

Comparando os métodos: analiticamente

- Regressão logística não é modelo generativo, pois não estima $X|Y$
- LDA é um caso particular de QDA, mas pode ter performance melhor
- Todo classificador com fronteiras de decisão linear é um naive Bayes (*e.g.*, LDA)
- Quando naive Bayes usa $f_{kj}(x_j)$ como $N(\mu_{kj}, \sigma_j^2)$, vira LDA com $\mathbf{\Sigma} = \text{diag}(\sigma_j^2)$
- QDA e naive Bayes não são casos particulares um do outro
  - QDA possui termos quadráticos (*e.g.*, $x_j x_k$)
  - Naive Bayes possui somas de funções sobre cada componente x_j
- Ou seja, métodos fazem hipóteses distintas; nenhum é uniformemente melhor

Comparando os métodos: empiricamente



Capítulo 4

- Classificação: setup geral
- Métodos discriminativos: regressão logística e regressão multinomial
- Métodos generativos: LDA, QDA e naive Bayes
- Avaliação de classificadores
- Modelos lineares generalizados

Avaliando métodos de classificação

- Em geral, para métodos de classificação, usamos a perda 0-1: $\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{[\hat{y}_i \neq y_i]}$
- Ela não é bem informativa: será que os erros são balanceados entre classes?
- Uma descrição melhor é a matriz de confusão:

		Actual class	
		Negative	Positive
Predicted class	Negative	True Negative (TN)	False Negative (FN) (or β)
	Positive	False Positive (FP) (or α)	True Positive (TP)

Exemplo: LDA

- Vamos usar LDA para prever **`default`** num dataset de 10 mil pessoas com previsão "sim" se $\mathbb{P}[\texttt{default} = \text{yes}|X] > 0.5$

		True default status	
		Negative	Positive
Predicted	Negative	9,644	252
	Positive	23	81

- A taxa de erro (de treino) é muito baixa: $(23 + 252/10000) = 2.75\%$
- Mas se prevermos sempre "não", o erro é de $(333/10000) \approx 3.33\%$
- Problema: a taxa de falsos negativos é alta: $252/333 \approx 76\%$
- Uma solução: mudar o threshold de 0.5 para 0.2

Exemplo: escolha de threshold

- Como a tabela anterior muda se mudarmos a nossa previsão de "sim" quando $\mathbb{P}[\texttt{default} = \text{yes}|X] > 0.2$ (versus 0.5)?

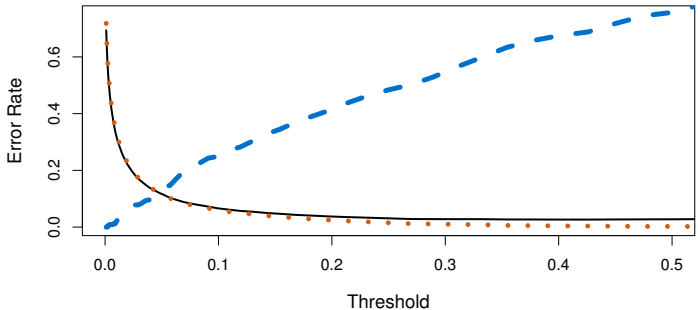
		True default status	
		Negative	Positive
Predicted	Negative	9,644	252
	Positive	23	81

	True default status	
	Negative	Positive
Negative	9,432	138
Positive	235	195

- A taxa de falso negativos agora é: $138/333 \approx 41\%$
- Mas a taxa de falso positivo subiu de $23/9667 \approx 0.2\%$ para $235/9667 \approx 2.4\%$
- E a taxa de erro (de treino) aumentou: $(235 + 138)/10000 \approx 3.7\%$
- Dependendo do problema, essa mudança pode ser útil (*e.g.*, em tribunais)

Exemplo: efeito do threshold no erro

- Vamos visualizar a dependência da taxa de erro no threshold:



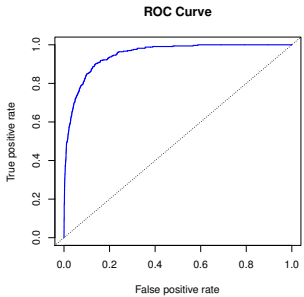
– – – Taxa de falso negativo (FN/P)

· · · · · Taxa de falso positivo (FP/N)

—— Perda 0-1 ou taxa de erro ((TP+TN)/(P+N))

Exemplo: curva ROC

- Curva ROC mostra a performance de um classificador para todos os thresholds



- A área sob a curva (AUC) mede a qualidade do classificador
  - Um classificador que sugere "não" com probabilidade 0.5 tem AUC de 0.5
  - Quanto mais perto de 1, melhor a AUC

Perguntas para revisão

- Por que não usamos métodos de regressão para problemas de classificação?
- Por que, mesmo querendo minimizar a perda 0-1, não é recomendado minimizá-la diretamente? O que fazer ao invés?
- Qual é a perda sendo minimizada pela regressão logística?
- Como encontrar o vetor de coeficientes $\hat{\beta}$ em regressão logística?
- Como interpretar os coeficientes de regressão logística?
- Como generalizar regressão logística para mais de dois casos $y \in \{0, 1\}$?
- O que são métodos generativos? Como diferenciar LDA, QDA e Naive Bayes?
- Sob que circunstâncias podemos encarar regressão logística, LDA, QDA e Naive Bayes como casos particulares uns dos outros? Quando preferir um ao outro?
- Como avaliar métodos de classificação? O que é curva ROC?