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 3

4 de março de 2024

Sumário

Notação

1. Matriz de dados $\mathbf{X}$, de ordem $n \times p$; n amostras (indivíduos), p variáveis.

$$\mathbf{X} = [x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}.$$

2. linhas de $\mathbf{X}$: $x_1, \ldots, x_n$; cada x_i é um vetor $p \times 1$;

   colunas de $\mathbf{X}$: $\mathbf{x}_1, \ldots, \mathbf{x}_p$; cada $\mathbf{x}_j$ é um vetor $n \times 1$.
3. Podemos escrever

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p] = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{bmatrix}.$$

4. y_i=i-ésima observação, $\mathbf{y} = (y_1, \ldots, y_n)^\top$. No caso de AE supervisionado, y_i é resposta aos preditores x_i, num problema de regressão, e é o rótulo da i-ésima classe, num problema de classificação.
   Dados: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, cada x_i um vetor $p \times 1$.

Tipos de dados

- grande número de amostras, n, e pequeno número de variáveis, p;
- pequeno número de amostras, n, e grande número de variáveis, p;
- grande número de amostras, n, e grande número de variáveis, p.
- Dados de Alta Dimensão: $n < p$.
- SAS (www.sas.comenus/insightsbig–datawhat–is–big–data.html): "Big data is a term that describes a large volume of data - both structured and unstructured - that inundates a business on a day-to-day basis. But it is not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves."

Dados de alta dimensão

- Alta dimensão relativa: modelos com muita variáveis (p) comparado com o número de amostras (n), mas usualmente com $p < n$;
- Alta dimensão moderada: modelos com número de variáveis proporcional ao número de amostras, usualmente $p > n$;
- Alta dimensão: modelos com mais variáveis do que amostras, e o número de variáveis cresce polinomialmente ou exponencialmente com n.

Tipos de dados

- Dados estruturados: informação organizada que se ajusta a estruturas usuais de bases de dados, relativamente fáceis de armazenar e analisar. Exemplos usuais de dados numéricos ou não, que podem ser dispostos em uma matriz de dados.
- Dados não estruturados: tudo que não se encaixa no item anterior, como arquivos de textos, páginas da *web*, email, midias sociais etc.
- Os quatro V's dos Big Data: VOLUME (escala dos dados); VARIEDADE (formas diferentes de dados); Velocidade (análise de *streaming data*); VERACIDADE (incerteza sobre os dados). (www.ibmbigdatahub.com)

Tipos de dados

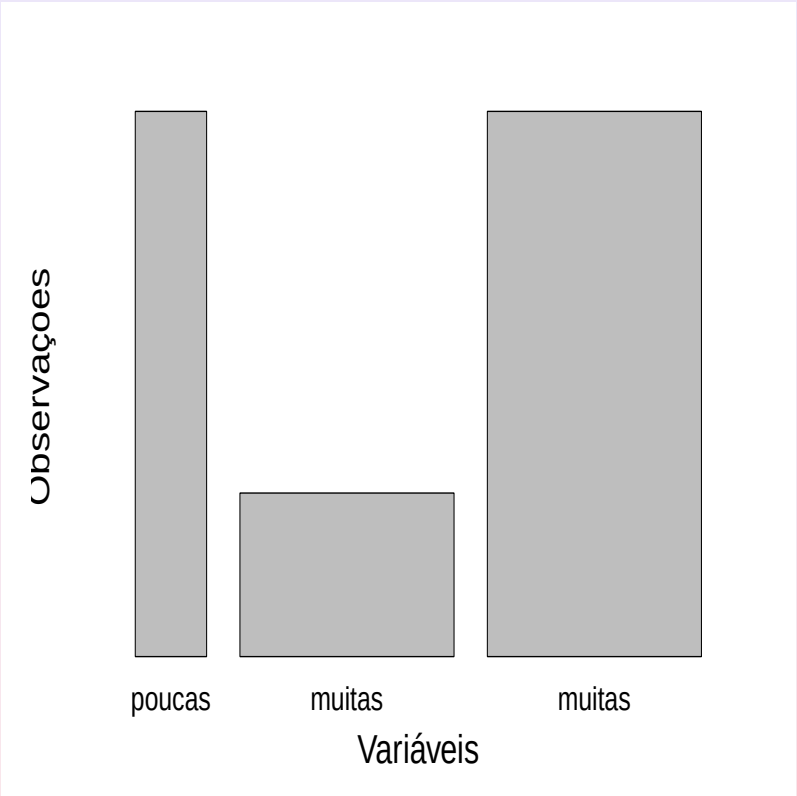

Figura 1: Tipos de dados

AE supervisionado

- Big Data implica em Big Models
- Big model: número grande de parâmetros (p) a serem estimados por algum método estatístico, comparado com o número de observações (n).
- Exemplo: regressão linear

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

- Problema se $p >> n$ (dados de alta dimensão)
- Estimação em modelos lineares com dados de alta dimensão pode ser tratada:
  (a) usando técnicas de redução da dimensão, como por exemplo ACP, AF e ACI;
  (b) usando estimação penalizada (regularização);
  (c) métodos bayesianos;
- No caso de modelos não lineares: árvores de decisão, redes neurais, bagging, boosting.

AE supervisionado: Exemplo 1

- Se quisermos saber se há relação entre o consumo privado (variável C) e renda disponível (variável Y) de indivíduos de uma população, podemos escolher uma amostra de n indivíduos dessa população e medir essas duas variáveis nesses indivíduos, obtendo-se o conjunto de dados $\{(Y_1, C_1), \ldots, (Y_n, C_n)\}$.
- Em Economia, sabe-se, desde Keynes, que o gasto com o consumo de pessoas (C) é uma função da renda pessoal disponível (Y), ou seja

$$C = f(Y),$$

para alguma função f.

- Para se ter uma ideia de como é a função f para essa comunidade, podemos construir um gráfico de dispersão entre Y e C. Com base em um conjunto de dados hipotéticos com $n = 20$, esse gráfico está apresentado na Figura 1 e é razoável postular o modelo

$$C_i = \alpha + \beta Y_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

em que (Y_i, C_i), $i = 1, \ldots, n$ são os valores de Y e C efetivamente observados e ε_i, $i = 1, \ldots, n$ são variáveis não observadas, chamadas **erros**. Aqui, $n = 20$ e $p = 1$.

AE supervisionado: Exemplo 1

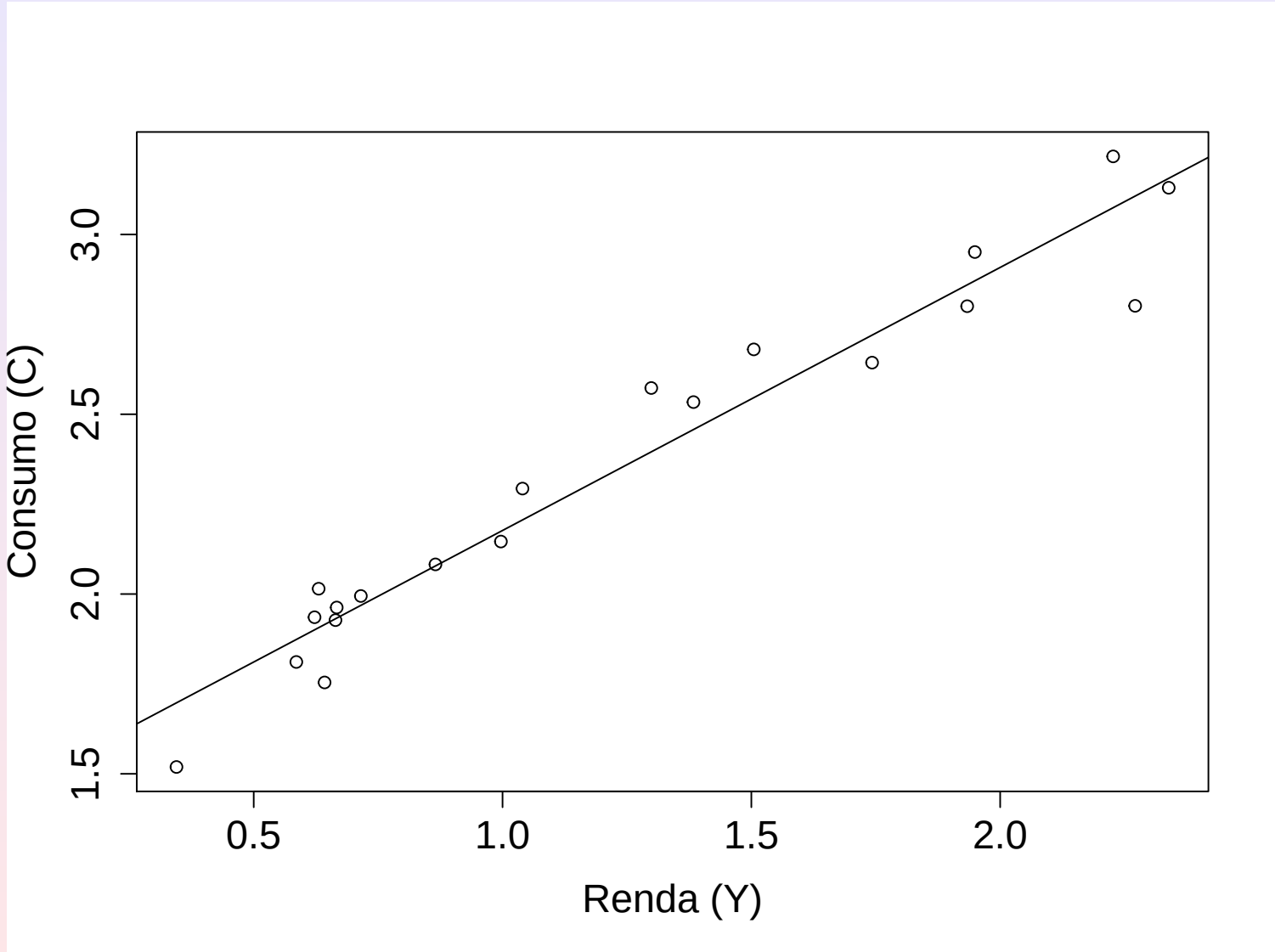

Figura 1: Relação entre renda e consumo de 20 indivíduos.

AE supervisionado: Exemplo 2

- Os dados apresentados na planilha **Esforço** são provenientes de um estudo sobre teste de esforço cardiopulmonar em pacientes com insuficiência cardíaca realizado no InCor da Faculdade de Medicina da USP. Um dos objetivos do estudo é comparar os grupos formados pelas diferentes etiologias quanto às respostas respiratórias e metabólicas obtidas do teste de esforço cardiopulmonar. Outro objetivo do estudo é saber se alguma das características observadas (ou combinação delas) pode ser utilizada como fator prognóstico de óbito.
- Nosso objetivo poderia ser desenvolver um modelo que possa ser usado para prever o consumo de oxigênio **VO2** (variável resposta, y) com as informações sobre **Carga** na esteira ergométrica (x_1) e **IMC** (x_2) (preditores). Um modelo de regressão linear seria

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, \ldots, 127. \tag{3}$$

Aqui, $p = 2$ e teremos ques estimar $\beta_0, \beta_1, \beta_2$ mais os parâmetros associados à variável ε_i. A Figura 2 mostra que o modelo (plano) não parece ser adequado, há vários pontos bastante afastados do plano.

AE supervisionado: Exemplo 2

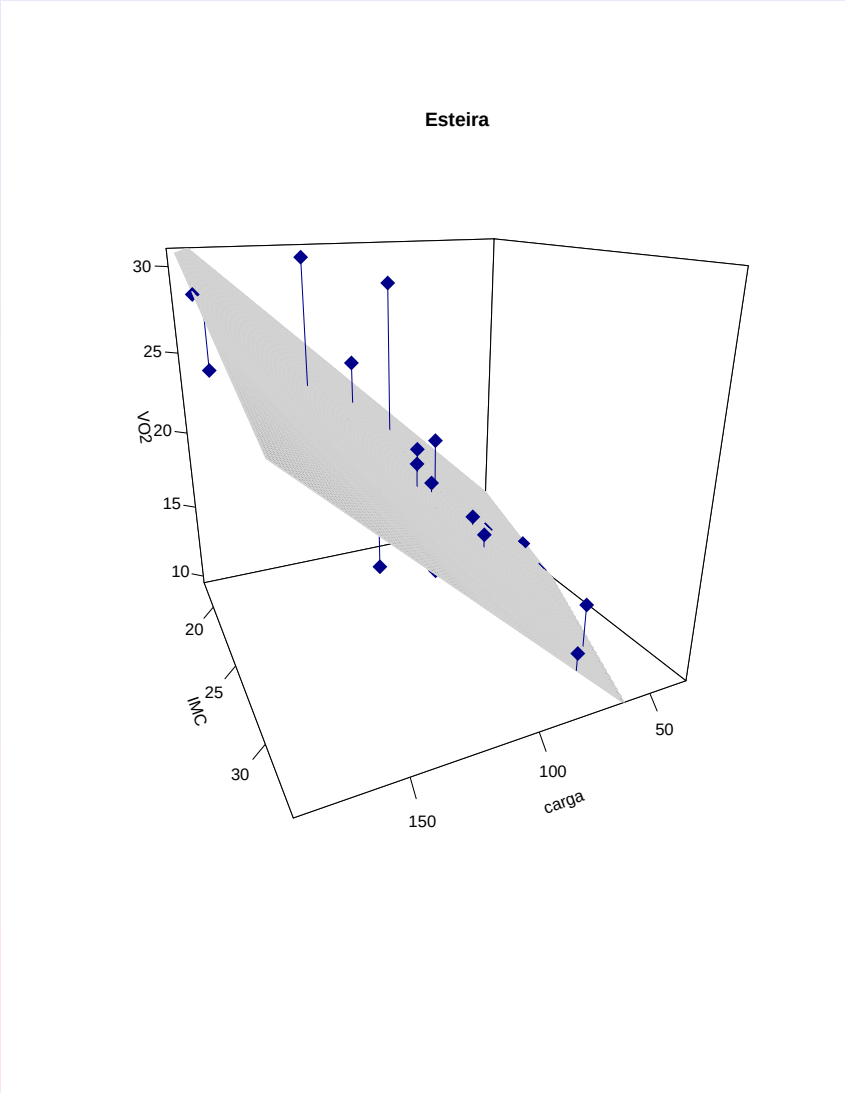

Figura 2: Consumo de oxigênio como função de Carga e IMC.

AE supervisonado: regressão

- Dados o vetor $\mathbf{y}$ de variáveis respostas e os preditores $\mathbf{x}_i$, o modelo geral é da forma

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4}$$

  com $E(\varepsilon_i) = 0$, ε_i ortogonal a $\mathbf{x}_i$ e f desconhecida, chamada de informação sistemática.
- O objetivo do AE é encontrar métodos para estimar f.
- Dois motivos para estimar f: Previsão e Inferência

Previsão

- Obtido o estimador $\hat{f}$, obtemos o previsor de $\mathbf{y}$

$$\hat{\mathbf{y}} = \hat{f}(\mathbf{X}).$$

- A acurácia de $\hat{\mathbf{y}}$ como previsor de $\mathbf{y}$ depende de dois erros:

  erro redutível: introduzido pelo estimador de f; assim chamado porque podemos melhorar a acurácia de $\hat{f}$ usando uma técnica de AE mais apropriada;

  erro irredutível: mesmo usando o melhor estimador de f, esse erro depende de ε, que não pode ser previsto usando $\mathbf{X}$.

- Supondo $\hat{f}$ e $\mathbf{X}$ fixos, pode-se ver que

$$\begin{aligned} E(\mathbf{y} - \hat{\mathbf{y}})^2 &= E[f(\mathbf{X}) + \varepsilon - \hat{f}(\mathbf{X})]^2 \\ &= E[f(\mathbf{X}) - \hat{f}(\mathbf{X})]^2 + \text{Var}(\varepsilon). \end{aligned} \tag{5}$$

O primeiro termo do segundo membro representa o erro redutível e, o segundo termo, o erro irredutível. O objetivo é minimizar o erro redutível.

Inferência

O interesse pode não ser fazer previsões, mas entender como a resposta é afetada pela variação dos preditores.
Interesse nos seguintes tópicos:

- Identificar alguns preditores importantes, dentre todos os preditores;
- Relação entre a resposta e cada um dos preditores; no Exemplo 2, y cresce com a carga, mas decresce com o IMC;
- A relação entre a resposta e cada preditor é linear, ou mais complicada? Modelos lineares fornecem interpretações mais simples, mas em geral previsões menos acuradas. Modelos não lineares fornecem previsões mais acuradas, mas o modelo perde em interpretabilidade.

Métodos paramétricos

- Fazemos alguma suposição sobre a forma de f, por exemplo, o modelo de regressão dado em (1). O problema simplifica-se, pois temos que estimar $p + 1$ parâmetros.
- Selecionado o modelo, temos que ajustá-lo aos dados de treinamento (*treinar* o modelo). No caso do modelo (1), o método mais usado é Mínimos Quadrados (MQ). Mas há outros métodos, como SVM.
- O ajuste do modelo (1) por MQ pode ser pobre, como no Exemplo 2 (ver Figura 2).
- Nesse caso, pode-se tentar modelos mais flexíveis, escolhendo-se outras formas funcionais para f, incluindo-se modelos não lineares.
- Modelos mais flexíveis envolvem estimar um número muito grande de parâmetros, aparecendo o problema do super-ajustamento (*overfitting*).

Métodos não paramétricos

- Nesse caso, não fazemos nenhuma hipótese sobre a forma funcional de f.
- Como o problema não se reduz a estimar um número pequeno de parâmetros, necessitaremos de um número grande de observações para obter estimadores acurados de f.
- Vários métodos podem ser usados:
  - ◇ usando kernels
  - ◇ usando polinômios locais (e.g Lowess)
  - ◇ usando splines
  - ◇ usando polinômios ortogonais (e.g. Chebyshev)
  - ◇ usando outras bases ortogonais (e.g. Fourier, ondaletas)

Acurácia e interpretação

- métodos menos flexíveis (e.g regressão linear) (ou mais restritivos) em geral são menos acurados e mais fáceis de interpretar.
- métodos mais flexíveis (e.g splines) são mais acurados e mais difíceis de interpretar.
- Para cada conjunto de dados, um método pode ser preferível a outros.
- Escolha do método é a parte mais difícil do AE/ML.

Qualidade do ajuste

- No caso de regressão, a medida de ajuste mais usada é o Erro Quadrático Médio (EQM), dado por

$$\text{EQM} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2, \tag{6}$$

onde $\hat{f}(x_i)$ é o preditor de y para a i-ésima observação.
- O EQM acima é calculado no conjunto de treinamento que produz $\hat{f}$. É chamado *EQM de treinamento*.
- Todavia, estamos mais interessados na acurácia do ajuste para os dados de teste.
- Se tivermos um grande número de dados de teste, poderemos calcular o *EQM de teste*,

$$\text{Média}(y_0 - \hat{f}(\mathbf{x}_0))^2, \tag{7}$$

que é o erro de previsão quadrático médio para as observações teste $(\mathbf{x}_0, y_0)$.
- Se não tivermos observações teste use (6). Na maioria dos casos o EQM de treinamento é menor que o EQM de teste.
- Para calcular o EQM de treinamento, usa-se CV.

Viés/variância

- Para um dado $(\mathbf{x}_0, y_0)$,

$$E[y_0 - \hat{f}(\mathbf{x}_0)]^2 = \text{Var}[\hat{f}(\mathbf{x}_0)] + [\text{Vies}(\hat{f}(\mathbf{x}_0))]^2 + \text{Var}(\varepsilon). \quad (8)$$

- Para minimizar (8), selecionamos um método que simultaneamente tiver baixo viés e baixa variância. O EQM de teste, em geral, apresenta uma forma de U, resultante da competição entre viés e variância.
- Métodos de AE mais flexíveis têm viés baixo e variância grande.
- Na prática, f não é conhecido e não é possível calcular o EQM de teste, viés e variância para um método de AE.

Classificação

- No caso de respostas $y_1, \ldots, y_n$ qualitativas temos um problema de **classificação**.
- Formalmente, seja $(\mathbf{x}, y)$, de modo que $\mathbf{x} \in \mathbb{R}^p$ e $y \in \{-1, 1\}$. Então, um **classificador** é uma função $g : \mathbb{R}^p \to \{-1, 1\}$ e a **função erro** ou **risco** é a probabilidade de erro, $L(g) = P\{g(X) \neq Y\}$.
- Obtendo-se um estimador de g, digamos $\hat{g}$, sua acurácia pode ser medida pelo estimador de $L(g)$, chamado de **taxa de erro de treinamento**, que é a proporção de erros gerados pela aplicação de $\hat{g}$ às observações de treinamento, ou seja,

$$\hat{L}(\hat{g}) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i), \tag{9}$$

sendo que $\hat{y}_i = \hat{g}(x_i)$ é o rótulo (-1 ou 1) da classe prevista usando $\hat{g}$.

Classificador de Bayes

- O interesse está na **taxa de erro de teste**

$$\text{Média}(I(y_0 \neq \hat{y_0}), \tag{10}$$

para observações de teste $(\mathbf{x}_0, y_0)$. Um bom classificador tem (10) pequeno.

- Pode-se provar que (10) é minimizado, em média, por um classificador que associa cada observação à classe mais provável, dados os preditores; ou seja, temos que maximizar

$$P(y = j|\mathbf{x} = \mathbf{x}_0). \tag{11}$$

Tal classificador é chamado de Bayes.

- No caso de duas classes, classificar na classe -1 se $P(y = -1|\mathbf{x} = \mathbf{x}_0) > 0,5$ e na classe 1 c.c. O classificador de Bayes produz a menor taxa de erro e será dada por $1 - \max_j P(y = j|\mathbf{x} = \mathbf{x}_0)$. A taxa de erro de Bayes global é dada por $1 - E(\max_j P(y = j|\mathbf{x} = \mathbf{x}_0))$, onde $E(\cdot)$ é caculada sobre todos os valores de $\mathbf{x}$.

K-ésimo vizinho mais próximo

- O classificador de Bayes não pode ser calculado na prática, pois não conhecemos a distribuição condicional de y dado $\mathbf{x}$.
- Uma possibilidade é estimar a distribuição condicional e, então, estimar (11).
- O classificador K-ésimo vizinho mais próximo (K-nearest neighbors, KNN) estima tal distribuição da seguinte maneira:
  - (i) Escolha $K > 0$ inteiro e uma observação teste $\mathbf{x}_0$.
  - (ii) O classificador KNN primeiro identifica os K pontos do conjunto de treinamento mais próximos de $\mathbf{x}_0$; chame-os de $\mathcal{N}$.
  - (iii) Estime a probabilidade condicional da classe j por

$$P(y = j|\mathbf{x} = \mathbf{x}_0) = \frac{1}{K} \sum_{i \in \mathcal{N}} I(y_i = j). \tag{12}$$

  - (iv) Classifique $\mathbf{x}_0$ na classe com a maior probabilidade condicional.
- Escolha de K crucial; resultado depende dessa escolha.

Modelos para séries temporais

(1) Consideremos uma série temporal multivariada $\mathbf{z}_t, t = 1, \ldots, T$ em que $\mathbf{z}_t$ contém valores de d variáveis e uma série temporal univariada Y_t, $t = 1, \ldots, T$. O objetivo é fazer previsões de Y_t, para horizontes $h = 1, \ldots, H$, com base nos valores passados de Y_t e de $\mathbf{z}_t$. Uma suposição básica é que o processo $\{Y_t, \mathbf{z}_t\}$, $t \geq 1$ seja estacionário fraco (ou de segunda ordem), veja Morettin (2017), com valores em $\mathbb{R}^{d+1}$. Para $p \geq 1$, consideramos o processo vetorial n-dimensional

$$\mathbf{x}_t = (Y_{t-1}, \ldots, Y_{t-p}, \mathbf{z}_t^\top, \ldots, \mathbf{z}_{t-r}^\top)^\top, \quad \text{com } n = p + d(r+1).$$

(2) O modelo a considerar é uma extensão do modelo (4), ou seja,

$$Y_t = f(\mathbf{x}_t) + e_t, \quad t = 1, \ldots, T. \tag{13}$$

O modelo geral

(3) Como em (4), f é uma função desconhecida e e_t tem média zero e variância finita. O objetivo é estimar f e usar o modelo para fazer previsões para um horizonte h por meio de

$$Y_{t+h} = f_h(\mathbf{x}_t) + e_{t+h}, \quad h = 1, \ldots, H, \ t = 1, \ldots, T.$$

Para uma avaliação do método de previsão, a acurácia é

$$\Delta_h(\mathbf{x}_t) = |\widehat{f}_h(\mathbf{x}_t) - f_h(\mathbf{x}_t)|.$$

(4) Em geral, a norma L_q é $E\{|\Delta_h(\mathbf{x}_t)|\}^q$, sendo que as mais comumente usadas consideram $q = 1$ (correspondente ao erro de previsão absoluto médio) ou $q = 2$ (correspondente ao errode previsão quadrático médio). A raiz quadrada dessas medidas também é utilizada.
Escolhendo-se uma função perda, o objetivo é selecionar f_h a partir de um conjunto de modelos que minimize o risco, ou seja, o valor esperado da norma L_q.

Modelos lineares

(1) Modelos frequentemente usados têm a forma (13) em que $f(\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$, com $\boldsymbol{\beta}$ denotando um vetor de $\mathbb{R}^n$. Estimadores de mínimos quadrados não são únicos se $n > T$. A ideia é considerar modelos lineares com alguma função de penalização, ou seja que minimizem

$$Q(\boldsymbol{\beta}) = \sum_{t=1}^{T-h} (Y_{t+h} - \boldsymbol{\beta}^\top \mathbf{x}_t)^2 + p(\boldsymbol{\beta}),$$

em que $p(\boldsymbol{\beta})$ depende, além de $\boldsymbol{\beta}$, de $\mathbf{Z}_t$, de um parâmetro de suavização λ e de eventuais hiperparâmetros.

(2) Este processo denomina-se **regularização**. Há várias formas de regularização como aquelas conhecidas por *Ridge*, *Lasso*, *Elastic Net* e generalizações. Detalhes sobre esses processos de regularização são apresentados no Capítulo 8.

Modelos não lineares

(1) Num contexto mais geral, o objetivo é minimizar

$$S(f) = \sum_{t=1}^{T-h} [Y_{t+h} - f(\mathbf{x}_t)]^2,$$

para f pertencendo a algum espaço de funções $\mathcal{H}$.

(2) Por exemplo, podemos tomar f como uma função contínua, com derivadas contínuas, ou f apresentando alguma forma de descontinuidade etc. Esses espaços, são, em geral, de dimensão infinita e a solução pode ser complicada. Para contornar esse problema, pode-se considerar uma coleção de espaços de dimensão finita $\mathcal{H}_d$, para $d = 1, 2, \ldots$, de tal sorte que $\mathcal{H}_d$ convirja para $\mathcal{H}$ segundo alguma norma. Esses espaços são denominados **espaços peneira** (*sieve spaces*).

(3) Para cada d, consideramos a aproximação

$$h_d(\mathbf{x}_t) = \sum_{j=1}^{J} \beta_j h_j(\mathbf{x}_t),$$

em que $h_j(\cdot)$ são **funções base** para $\mathcal{H}_d$ e tanto J como d são funções de T. Podemos usar *splines*, polinômios, funções trigonométricas, ondaletas etc. como funções base.

Modelos não lineares

(4) Se as funções base são conhecidas, elas são chamadas *linear sieves* e se elas dependem de parâmetros a estimar, são chamadas *nonlinear sieves*.

(5) Exemplos de non linear sieves são as árvores de decisão e as redes neurais.

(6) Métodos em AE ou ML são dedicados a observações de variáveis independentes e identicamente distribuídas. O caso de séries temporais é mais complicado e foge ao escopo deste curso. Apresentaremos apenas algumas ideias relacionadas a esse tópico.

Referências

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

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