

Estudo do livro: An Introduction to Statistical Learning

Capítulos 1 e 2

Agatha Rodrigues

Departamento de Estatística da UFES e R-Ladies Vitória
setembro/2020

Agatha Rodrigues



- Professora do Departamento de Estatística da UFES.
- Coordenadora do DaSLab - Laboratório de Data Science da UFES -
<https://daslab-ufes.github.io/>



- Co-organizadora da R-Ladies Vitória ❤️



- Currículo Lattes - <http://lattes.cnpq.br/3445977720574534>
- GitHub:[agathasr](#)
- Anteriormente:
 - Doutora em Estatística - IME/USP
 - Mestre em Estatística - IME/USP
 - Bacharel em Estatística - UFSCar



R-Ladies?



- R-Ladies é uma organização mundial que **promove a diversidade de gênero** na comunidade R.
- Capacitar pessoas de gêneros sub-representados, criando e fortalecendo redes colaborativas dentro da comunidade R para que elas alcancem todas e quaisquer funções e áreas de participação no mundo da tecnologia.

Como

- Promovendo meetups (encontros) e mentorias.
- Garantindo espaço amigável e seguro.



O Capítulo da cidade de Vitória foi criado em 29 de setembro de 2019.

- Código de conduta - R-Ladies
- Saiba mais:
 - RLadies Global: <https://rladies.org/>
 - MeetUp: <https://www.meetup.com/pt-BR/R-Ladies-Vitoria>
 - Twitter: @RLadiesGlobal, @rladiesvix
 - Instagram: @rladiesvix
 - Github: https://github.com/rjladies/meetup-presentations_vitoria



Vamos começar?

Material

Livro principal

Livro [An Introduction to Statistical Learning - with Applications in R \(ISLR\)](#), de Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani.

Gareth James:

Data is the sword of the 21st century, those who wield it well, the Samurai.

Outras referências

- Livro [Introdução à Ciência de Dados Fundamentos e Aplicações](#), de Pedro Morettin e Julio Singer.
- Livro [Machine Learning sob a ótica estatística: Uma abordagem preditivista para estatística com exemplos em R](#), com Tiago Mendonça e Rafael Izbicki.

Sobre esse minicurso

- **Objetivo:** Estudar os conceitos do livro ISLR.

O que é aprendizado estatístico? E Estatística?

Sobre o livro ISLR

Um pouco de Estatística Básica

Notação e conceitos

Organização do livro

O que é
aprendizado
estatístico?

Do livro ISLR

O aprendizado estatístico (AE) se refere a um vasto conjunto de ferramentas para a compreensão de dados.

Do livro de Morettin e Singer

O AE consiste na utilização de modelos estatísticos acoplados a algoritmos computacionais desenvolvidos para extrair informação de conjuntos de dados contendo, em geral, muitas unidades amostrais e muitas variáveis.

E Estatística?

Wikipédia

Estatística é a ciência que utiliza-se das teorias probabilísticas para explicar a frequência da ocorrência de eventos, modelar a aleatoriedade e a incerteza de forma a estimar ou possibilitar a previsão de fenômenos futuros.

Guia do Estudante

É a ciência que analisa e interpreta dados no estudo de fenômenos naturais, econômicos e sociais.

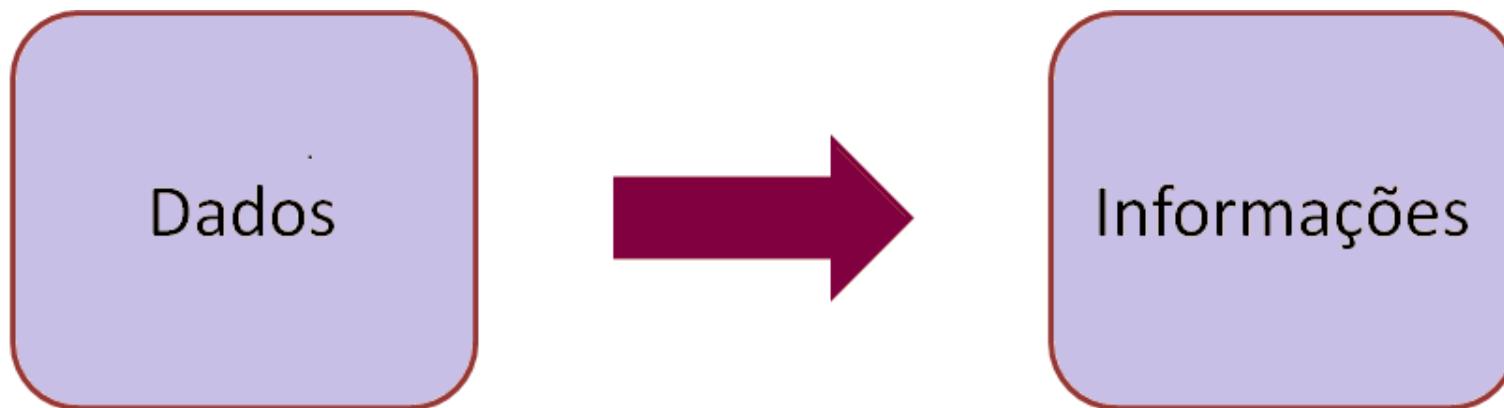
W. Allen Wallis

A Estatística pode ser definida como um conjunto de métodos utilizados para tomada de decisões sábias em face à incerteza.

Carlos Alberto de Bragança Pereira

O Estatístico é um bruxo que faz afirmações científicas sobre estados e quantidades invisíveis. No entanto, diferente dos demais bruxos, ele atribui uma medida de incerteza as suas declarações.

Estatística



Algumas frases
que às vezes
escutamos

"Não precisa de Estatística para a Ciência de Dados"



Robert Tibshirani:

Statistics is the key of Data Science.

"Estatística só lida com pequenos volumes de dados (*small data*)"

Morettin e Singer:

Se uma das principais características da Ciência de Dados é analisar grandes conjuntos de dados (megadados), há mais de 200 anos os estatísticos têm se preocupado com a análise de vastos conjuntos de dados provenientes de censos, coleta de informações meteorológicas, observação de séries de índices financeiros etc., que tem essa característica.

Talvez seja o aspecto computacional necessário para lidar com megadados o que mascara os demais componentes daquilo que se entende por Ciência de Dados, pois em muitos casos, o interesse é dirigido apenas para o desenvolvimento de algoritmos cuja finalidade é aprender a partir dos dados, omitindo-se características estatísticas.

Sobre o livro ISLR

As premissas do livro

1. **Muitos métodos de aprendizado estatístico são relevantes e úteis em uma gama de disciplinas acadêmicas e não acadêmicas.** Ao invés de tentar considerar todas as possíveis abordagens (uma tarefa impossível), o livro é concentrado em apresentar os métodos que os autores acreditam serem mais amplamente aplicáveis.
2. **O aprendizado estatístico não deve ser visto como uma série de caixas pretas.** Portanto, os autores tentam descrever cuidadosamente o modelo, intuição, suposições, e compensações por trás de cada um dos métodos considerados.
3. **Embora seja importante saber qual trabalho é executado por cada engrenagem, não é necessário ter as habilidades para construir a máquina dentro da caixa.** Assim, a discussão de detalhes técnicos relacionados aos procedimentos de ajuste e propriedades teóricas é minimizada. Os autores presumem que o leitor se sinta confortável com os conceitos matemáticos básicos, mas não assumem uma pós-graduação em Matemática.
4. **Presumem que o leitor esteja interessado em aplicar métodos de aprendizagem estatística a problemas do mundo real.** Para facilitar isso, bem como para motivar as técnicas discutidas, é dedicada uma seção em cada capítulo aos laboratórios sobre o software R.

Pacote para os dados utilizados no livro

O pacote *ISLR* disponível no site do livro contém uma série de conjuntos de dados associados a esse livro.

Site do livro

O site do livro está localizado em

www.StatLearning.com

As bases de dados do livro

Nome	Descrição
Auto	Milhagem a gás, potência e outras informações para carros.
Boston	Valores de casas e outras informações sobre os subúrbios de Boston.
Caravan	Informações sobre pessoas que oferecem seguro de caravana.
Carseats	Informações sobre vendas de assentos de carro em 400 lojas.
College	Características demográficas, taxas de matrícula e muito mais para faculdades dos EUA.
Default	Registros de inadimplência de clientes para uma empresa de cartão de crédito.
Hitters	Recordes e salários de jogadores de beisebol.
Khan	Medições de expressão gênica para quatro tipos de câncer.
NCI60	Medições de expressão gênica para 64 linhas celulares de câncer.
OJ	Informações de vendas de suco de laranja Citrus Hill e Minute Maid.
Portfolio	Valores passados de ativos financeiros para uso na alocação de carteiras.
Smarket	Retornos percentuais diários para S&P 500 ao longo de um período de 5 anos.
USArrests	Estatísticas criminais por 100.000 habitantes em 50 estados dos EUA.
Wage	Dados de pesquisa de renda de homens na região do Atlântico central dos EUA.
Weekly	1.089 retornos semanais do mercado de ações por 21 anos.

Todos os conjuntos de dados estão disponíveis no pacote *ISLR*, com exceção de **Boston** (no pacote *MASS*) e **USArrests** (base do R).

Quem deveria ler esse livro?

- Este livro é destinado a qualquer pessoa interessada em usar métodos estatísticos modernos para modelagem e predição de dados.
- Este grupo inclui cientistas, engenheiros, analistas de dados, mas também indivíduos menos técnicos com graduação em campos não quantitativos, como ciências sociais ou negócios.
- O nível matemático deste livro é modesto e um conhecimento detalhado sobre operações de matrizes não é necessário.
- **É esperado que o leitor tenha pelo menos um conhecimento básico de Estatística.**



Um pouco de Estatística Básica

Vamos para a apresentação de Estatística Básica...

Exercício

Estudo de:

- Medidas descritivas;
- Alguns gráficos;
- Instalação do R e RStudio.

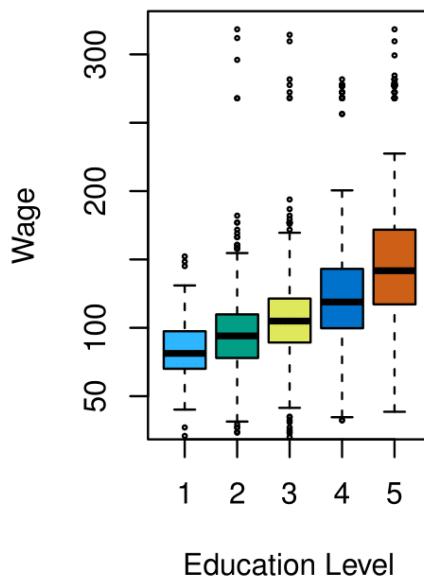
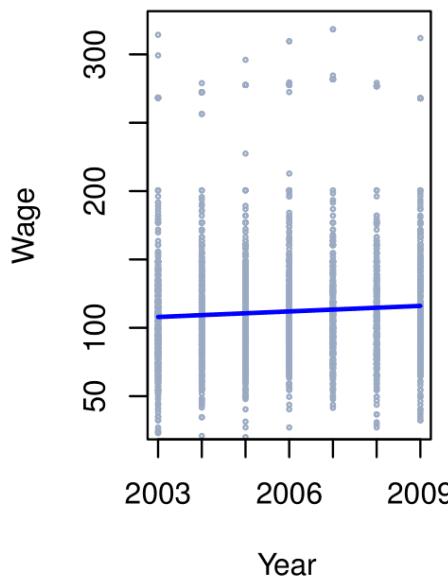
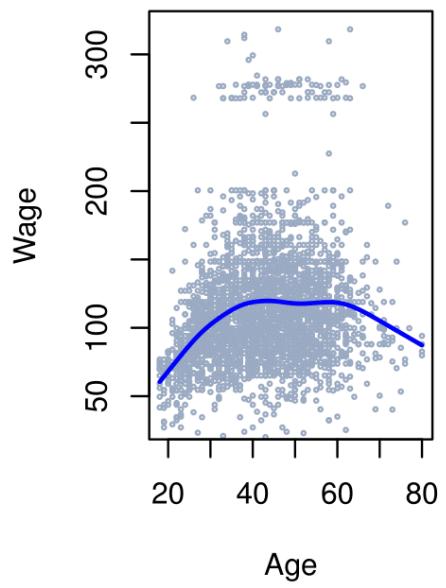
Link para materiais

<https://daslab-ufes.github.io/materiais>

Notação

Wage dataset

- Vamos avaliar uma série de fatores relacionados aos salários de um grupo de homens da região do atlântico dos Estados Unidos.
- Em particular, queremos entender a associação entre o **salário** (*wage*) com as seguintes variáveis:
 - **idade** (*age*);
 - **escolaridade** (*education level*);
 - **ano** (*year*);
 - **estado civil** (*marital*);
 - **etnia** (*race*);
 - **classificação do trabalho** (*jobclass*);
 - **indicador de saúde** (*health*);
 - **se tem seguro saúde** (*health_ins*).



Notação

O salário é a variável de saída (*output*), enquanto as demais variáveis são variáveis de entrada (*input*).

Sobre as variáveis de entrada

- **Sinônimos:** *input*, preditores, variáveis independentes, *features*, covariáveis, variável exógena.
- **Notação:** tipicamente denotadas pelo símbolo X , com um subscrito para distingui-las.
- Na motivação **Wage**, X_1 pode ser a idade, X_2 escolaridade, X_3 o ano, etc.

Sobre variável de saída

- **Sinônimos:** *output*, variável resposta, variável dependente, variáveis endógenas.
- **Notação:** tipicamente denotadas pelo símbolo Y .
- Na motivação **Wage**, a variável salário é denotada por Y .

Notação

- n denotará o número de observações;
- p para indicar o número de variáveis de entrada.

Por exemplo: o conjunto de dados **Wage** consiste em 8 variáveis de entrada para 3000 observações.

Portanto, temos $n = 3000$ e $p = 8$.

Notação

- Ao longo do livro ISLR, o índice i será usado para as observações (de 1 a n) e o índice j será usado as variáveis (de 1 a p).
- Seja x_{ij} o valor da j -ésima variável para a i -ésima observação, em que $i = 1, \dots, n$ e $j = 1, \dots, p$.
- Seja x_i um vetor de tamanho p , contendo as medidas das p variáveis para a i -ésima observação. Isso é,

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}.$$

- **Por exemplo:** na base de dados **Wage**, x_i é um vetor de tamanho p , consistindo nos valores das variáveis de entrada para o i -ésimo indivíduo, com $i = 1, \dots, n$.

Notação

- Seja y_i o valor da i -ésima observação da variável de saída, como o salário na base de dados **Wage**.
- Assim, os dados observados consistem em $x_1 \ y_1 \quad x_2 \ y_2 \quad \dots \ x_n \ y_n$, onde cada x_i é um vetor de comprimento p .

Se $p = 1$, então x_i é simplesmente um escalar.

Modelo

- Suponha que observamos uma resposta quantitativa Y e p diferentes preditores X_1, X_2, \dots, X_p .
- Presumimos que haja alguma relação entre Y e X_1, X_2, \dots, X_p , que pode ser escrita na forma geral

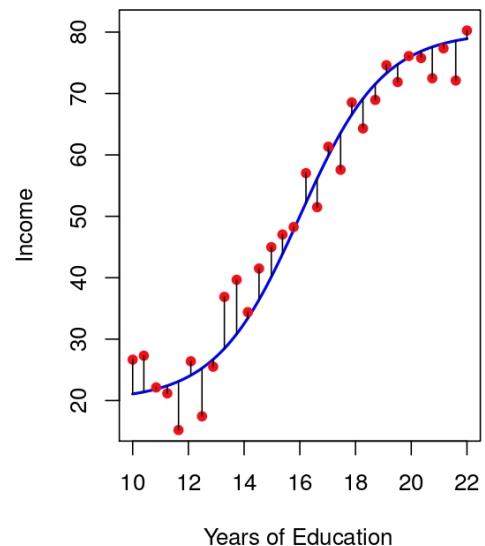
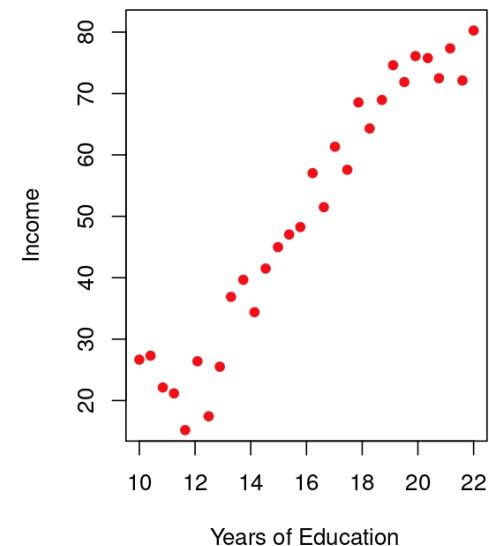
$$Y = f(X) + \epsilon$$

f é alguma função fixa e desconhecida de X_1, X_2, \dots, X_p ;

ϵ é um termo de erro aleatório e independente de X .

- f representa a informação que X fornece sobre Y .

Um exemplo simulado: renda versus educação



- Em essência, o aprendizado estatístico se refere a um conjunto de abordagens para estimar f .
- Vamos discutir alguns conceitos:
 - Por que estimar f ?
 - Como estimar f ?
 - Acurácia preditiva versus interpretabilidade do modelo.
 - Tipos de aprendizado estatístico.
 - Problema de regressão versus problema de classificação.
 - Como escolher o método?



Por que estimar f ?

Por que estimar f ?

- Existem duas razões principais pelas quais podemos desejar estimar f : **predição** e **inferência**.

Predição

Em muitas situações, um conjunto de entradas X está prontamente disponível, mas a saída não pode ser obtida facilmente. Podemos prever Y usando

$$\hat{Y} \quad \hat{f} \quad X$$

em que:

- \hat{f} representa nossa estimativa para f ;
- \hat{Y} representa a predição resultante para Y .

Tipos de erros

A acurácia de \hat{Y} como predição de Y depende dos seguintes dois erros:

- **erro redutível:** introduzido pela estimativa de f ; assim chamado porque podemos melhorar a acurácia de \hat{f} usando técnicas de AE mais apropriadas.

No entanto, mesmo que fosse possível formar uma estimativa perfeita para f , de forma que nossa resposta estimada assumisse a forma $\hat{Y} = f(X)$, nossa predição ainda teria algum erro nela! Isso ocorre porque Y também é uma função de ϵ , que não pode ser prevista usando X .

Portanto, a variabilidade associada a ϵ também afeta a acurácia de nossas predições.

- **erro irredutível:** não importa o quão bem estimemos f , não podemos reduzir o erro introduzido por ϵ .

Inferência

- Desejamos estimar f , mas agora com o interesse é entender a maneira como Y é afetado com as mudanças em X X_p .
- Podemos responder as seguintes questões:
 - Quais preditores estão associados à resposta?
 - Qual é a relação entre a resposta e cada preditor?
 - A relação entre Y e cada preditor pode ser resumida adequadamente usando uma equação linear ou a relação é mais complicada?

No exemplo Wage

- Quais variáveis de entrada contribuem para o salário?
- Quais geram o maior impulso nos salários?
- Quanto aumento na idade, por exemplo, está associado a um determinado aumento no salário?



Como podemos
estimar f ?

Como podemos estimar f ?

- Considere um conjunto de n observações.
- Essas observações são chamadas de **dados de treinamento** porque usaremos essas observações para treinar, ou ensinar, nosso método como estimar f .
- Seja x_{ij} o valor do j -ésimo preditor para a observação i , em que $i = 1, \dots, n$ e $j = 1, \dots, p$. Correspondentemente, seja y_i a variável de resposta para a i -ésima observação.
- Os dados de treinamento consistem em $x_1 \ y_1 \ x_2 \ y_2 \ \dots \ x_n \ y_n$, onde $x_i = [x_{i1} \ x_{i2} \ \dots \ x_{ip}]^T$.
- Nosso **objetivo** é aplicar um método de aprendizagem estatística aos dados de treinamento para estimar a função desconhecida f . Em outras palavras, queremos encontrar uma função \hat{f} tal que $Y \approx \hat{f}(X)$ para qualquer observação (X, Y) .
- Em termos gerais, a maioria dos métodos de aprendizagem estatística para essa tarefa pode ser caracterizada como **paramétrica** ou **não paramétrica**.

Métodos paramétricos

- Fazemos uma **suposição** sobre a forma funcional de f .
- Por exemplo, uma suposição muito simples é que f é linear em X :

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

Para estimar $f(X)$, basta estimar os p coeficientes $\beta_0, \beta_1, \dots, \beta_p$.

- A **desvantagem** potencial de uma abordagem paramétrica é que o modelo que escolhemos geralmente não corresponderá à verdadeira forma desconhecida de f . Se o modelo escolhido estiver muito longe de f verdadeiro, nossa estimativa será ruim.
- **Possível solução:** escolher modelos flexíveis que possam ajustar muitas formas funcionais possíveis para f .
- Em geral, ajustar um modelo mais flexível requer estimar um maior número de parâmetros.

Métodos não paramétricos

- Nos métodos não paramétricos, não fazemos suposições sobre a forma funcional de f . Em vez disso, eles buscam uma estimativa de f que chegue o mais próximo possível dos pontos de dados.
- Essas abordagens podem ter uma grande vantagem sobre as abordagens paramétricas: ao evitar a suposição de uma forma funcional específica para f
- As abordagens não paramétricas sofrem de uma grande desvantagem: uma vez que não reduzem o problema de estimar f a um pequeno número de parâmetros, um número muito grande de observações (muito mais do que normalmente é necessário para uma abordagem paramétrica) é necessário para obter estimadores acurados f .

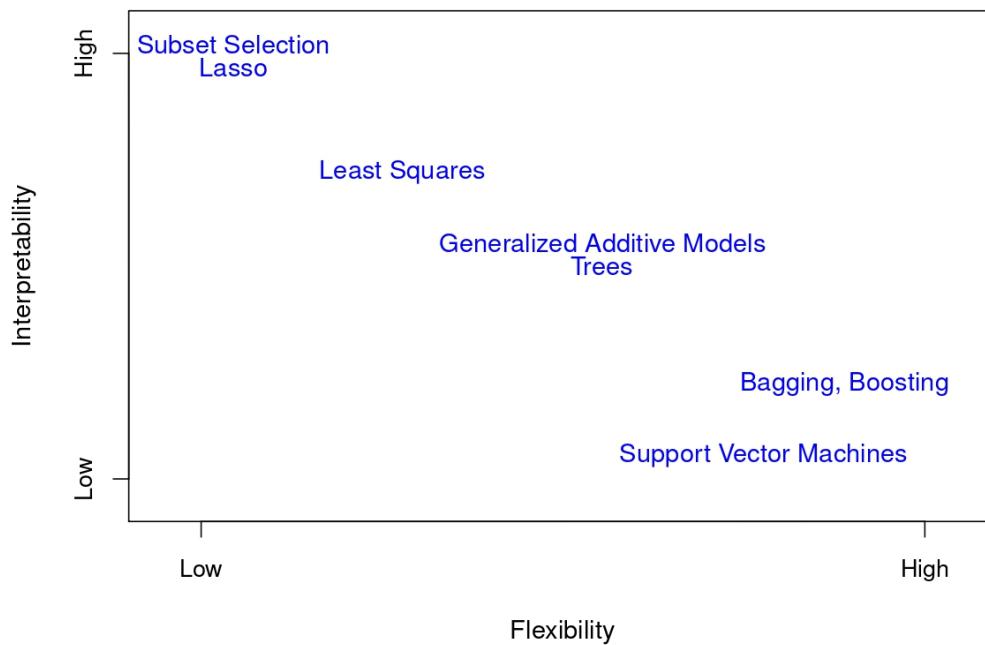
Acurácia preditiva versus interpretabilidade

O *trade-off* entre acurácia na predição e interpretabilidade do modelo

- Alguns métodos são menos flexíveis, ou mais restritivos, no sentido de que podem produzir apenas uma gama relativamente pequena de formas para estimar f .
- **Por que escolheríamos usar um método mais restritivo em vez de uma abordagem mais flexível?**

Há algumas razões. Se estamos interessados em inferência, os modelos restritivos são mais interpretáveis.

- Modelo interpretável: modelo de regressão linear.
- Abordagens muito flexíveis, como os splines discutidos no Capítulo 7, e os métodos de reforço discutidos no Capítulo 8, podem levar a estimativas tão complicadas de f que é difícil entender como qualquer preditor individual está associado a a resposta.



Mais sobre o tema:

Post: Interpretabilidade em modelos preditivos – discussões iniciais na área da saúde

<https://daslab-ufes.github.io/interpretability-in-predictive-model>

Tipos de Aprendizado

Tipos de Aprendizado

O AE pode ser supervisionado ou não supervisionado.

Supervisionado

- Envolve a construção de um modelo estatístico para prever ou estimar uma saída com base em uma ou mais entradas.
- Para cada observação do(s) preditor(es) $x_i, i = n$ há uma medida de resposta associada y_i .
- Queremos ajustar um modelo que relate a resposta aos preditores, com o objetivo de prever com a resposta para observações futuras (predição) ou melhor compreender a relação entre a resposta e os preditores (inferência).
- Com exceção do Cap. 10, todos os métodos do livro ISLR são para AE supervisionados.
- Exemplo de supervisionado: **Wage** dataset.

Não supervisionado

- Para cada observação $i \in [n]$, observamos um vetor de medidas x_i , mas nenhuma resposta associada y_i .
- A situação é chamada de não supervisionada porque não temos uma variável de resposta que possa supervisionar nossa análise.
- No livro: Capítulo 10.
- Podemos ter interesse em entender a relação entre variáveis ou entre observações.

Exemplo de não supervisionado: segmentação de marketing

- Podemos observar múltiplas características (variáveis) para **clientes em potencial**, como CEP, renda familiar e hábitos de compra.
- Acreditamos que os clientes se enquadram em grupos diferentes, como grandes compradores versus pequenos compradores.
- Se as informações sobre os padrões de gastos de cada cliente estivessem disponíveis, uma análise supervisionada seria possível. No entanto, essas informações não estão disponíveis, ou seja, não sabemos se cada cliente potencial gasta muito ou não.
- Neste cenário, podemos tentar agrupar os clientes com base nas variáveis medidas, de forma a identificar grupos distintos de potenciais clientes. Identificar esses grupos pode ser interessante porque pode ser que os grupos difiram em relação a alguma propriedade de interesse, como hábitos de consumo.

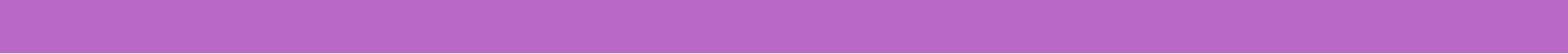
Regressão versus classificação

Problemas de regressão versus de classificação

- As variáveis podem ser caracterizadas como quantitativas ou qualitativas (também conhecidas como categóricas).
- Exemplos:
 - **Quantitativas** - idade, altura ou renda de uma pessoa, o valor de uma casa, e o preço de uma ação.
 - **Qualitativas** - a marca do produto comprado (marca A, B ou C), se uma pessoa é inadimplente (sim ou não), ou um diagnóstico de câncer (Leucemia Mielóide Aguda, Aguda Leucemia linfoblástica ou sem leucemia).

Em AE supervisionado

- Problema de **regressão**: Quando Y é uma variável quantitativa;
- Problema de **classificação**: Quando Y é uma variável qualitativa.



Como escolher o
método?

Avaliação da precisão do modelo

- **Pergunta interessante:** Por que precisamos estudar tantos métodos?

Não há almoço grátis: nenhum método domina todos os outros sobre todos os conjuntos de dados possíveis.

- **Objetivo:** Escolher qual método produz os melhores resultados dentre um conjunto de métodos candidatos.
- Para avaliar o desempenho de um método de AE em um determinado conjunto de dados, precisamos de **alguma forma para medir** o quanto bem suas previsões realmente correspondem aos dados observados.

Erro quadrático médio (EQM)

Em problemas de regressão, a medida mais comumente usada é o erro quadrático médio (sigla em inglês: MSE), dado por:

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

em que $\hat{f}(x_i)$ é a predição que \hat{f} fornece para a i -ésima observação.

- O EQM será pequeno se as respostas preditas forem muito próximas das respostas verdadeiras, e será grande se as respostas preditas e verdadeiras diferirem substancialmente.
- Não queremos saber o quanto bem o método funciona nos dados de treinamento. Em vez disso, estamos interessados na precisão das previsões que obtemos quando aplicamos nosso método a dados de teste **nunca vistos antes**.

Amostras treinamento e teste

- **Dados treinamento:** para o ajuste do método de AE

Ajustamos nosso método de AE em nossas observações de treinamento $x_1 \ y_1 \ x_2 \ y_2 \dots x_n \ y_n$, e obtemos a estimativa \hat{f} . Podemos então calcular $\hat{f}(x_1) \ \hat{f}(x_2) \dots \hat{f}(x_n)$.

- **Dados teste:** para a avaliação do ajuste

Sejam $x \ y$ observações não consideradas para treinar o método de AE.

- Queremos saber se $\hat{f}(x)$ é aproximadamente igual a y .

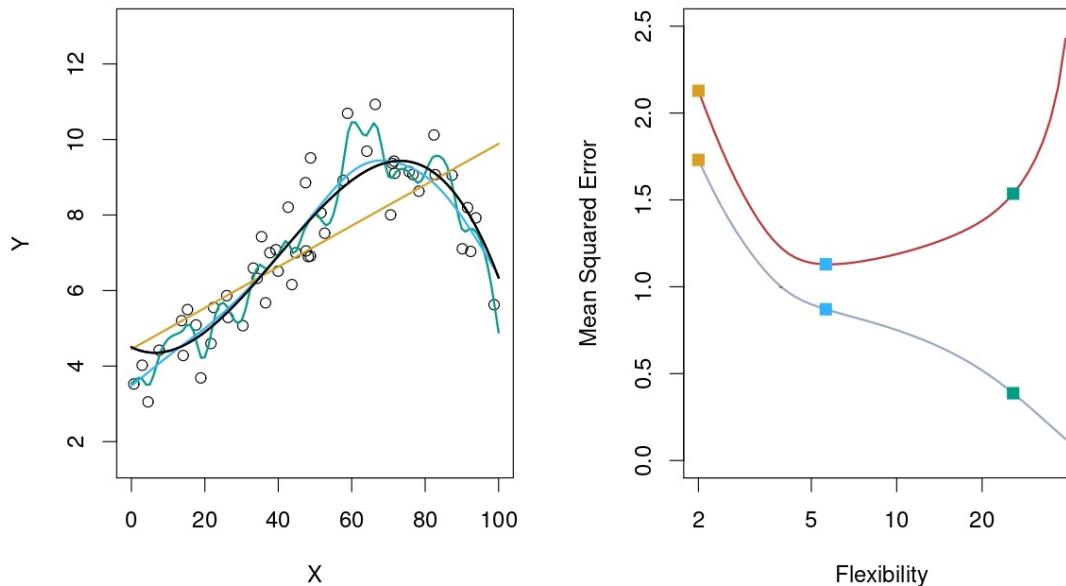
Escolha do método

- Uma medida de avaliação do ajuste é o cálculo do EQM na amostra teste.

Dentre um conjunto de métodos de AE, escolhemos aquele com o menor EQM da amostra teste.

- Mas daí surge uma pergunta: O método com menor EQM no treinamento também apresenta o menor EQM no teste?

Não necessariamente!



À esquerda: Dados simulados de f , mostrados em preto. Três estimativas de f : a linha de regressão linear (curva laranja) e dois ajustes de spline (curvas azul e verde). À direita: EQM de treinamento (curva cinza) e EQM de teste (curva vermelha). Os quadrados representam os EQMs de treinamento e teste para os três ajustes mostrados no painel esquerdo.

Em cenários de classificação

- Queremos estimar f com base nas observações de treinamento $x \ y \ x_n \ y_n$, mas agora $y \ y_n$ são observações qualitativas.
- A **taxa de erro de treinamento** é dada por:

$$\frac{1}{n} \sum_i^n I(y_i \neq \hat{y}_i)$$

em que \hat{y}_i é a classificação predita para i -ésima observação usando f e I é uma variável indicadora que vale 1 se $y_i \neq \hat{y}_i$ e 0, caso contrário.

- A **taxa de erro de teste** é calculada de maneira análoga, mas com a amostra test $x \ y$.



O que vem por aí

Organização do livro

- O **Capítulo 2** apresenta a terminologia e os conceitos básicos por trás do aprendizado estatístico.
- Os **capítulos 3 e 4** cobrem métodos lineares clássicos para regressão e classificação.
- Um problema central em todas as situações de aprendizado estatístico envolve a escolha do melhor método para uma determinada aplicação. Portanto, no **Capítulo 5** é introduzido validação cruzada e o bootstrap, que pode ser usado para estimar a precisão de vários métodos diferentes a fim de escolher o melhor.
- No **Capítulo 6**, é considerada uma série de métodos lineares que oferecem melhorias potenciais em relação à regressão linear padrão (Capítulo 3).

Organização do livro

- No **Capítulo 7** uma série de métodos não lineares que funcionam bem para problemas com uma única variável de entrada. Em seguida, mostramos como esses métodos podem ser usados para ajustar modelos aditivos não lineares para os quais há mais de uma entrada.
- No **Capítulo 8**, métodos baseados em árvores são investigados, incluindo *bagging*, *boosting* e *random forests*.
- Métodos de vetores de suporte (*support vector machine*) são discutidos no **Capítulo 9**.
- Finalmente, no **Capítulo 10**, é considerado uma configuração em que há variáveis de entrada, mas nenhuma variável de saída.

Obrigada!

- Currículo Lattes - <http://lattes.cnpq.br/3445977720574534>
- GitHub - [agathasr](#)
- Email - agatha.srodrigues@gmail.com