

Aprendizagem de Máquina

Conceitos Fundamentais

Juliana Freitas Pires

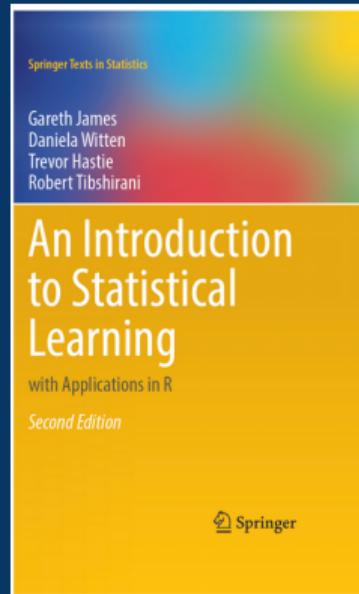
juliana.freitas@academico.ufpb.br

www.de.ufpb.br

UFPB

 Departamento de
ESTATÍSTICA

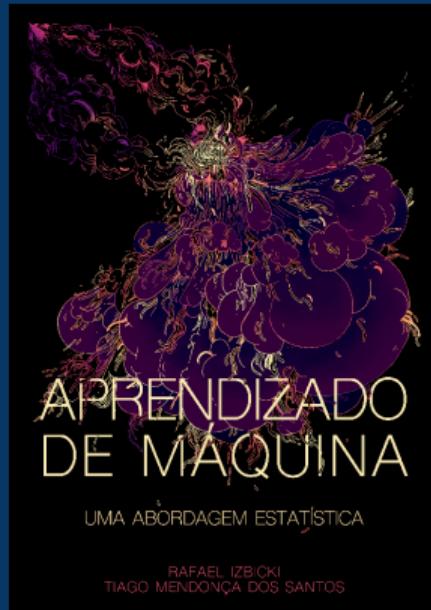
Bibliografia



- G. James, D. Witten, T. Hastie and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. 2nd ed., Springer, 2021.
(Download gratuito em: <https://www.statlearning.com/>).



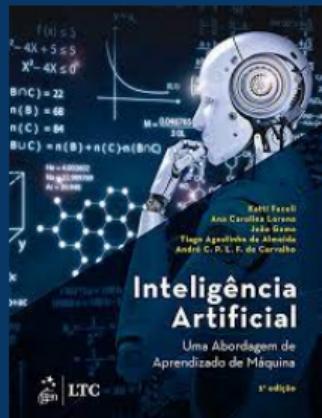
Bibliografia



- Izbicki, R. e Santos, T. M. dos. Aprendizado de máquina: uma abordagem estatística. 1^a edição. 2020.
(Download gratuito em: <http://www.rizbicki.ufscar.br/AME.pdf>).



Bibliografia



- FACELI, Katti; LORENA, Ana C.; GAMA, João; AL, et. Inteligência Artificial - Uma Abordagem de Aprendizado de Máquina. Grupo GEN, 2021. E-book. ISBN 9788521637509.

[https://integrada\[minhabiblioteca\].com.br/#/books/9788521637509/](https://integrada[minhabiblioteca].com.br/#/books/9788521637509/)



Introdução

Há algumas décadas, a maioria das soluções computacionais eram baseadas em programas que codificavam, passo a passo, as ações necessárias para resolver um dado problema.

Ainda que muito útil em determinados contextos, essa abordagem não era adequada para resolver vários problemas que seres humanos resolvem com facilidade, como:

- ▶ Reconhecer uma pessoa pelo seu rosto ou voz. Nós reconhecemos faces independentemente das expressões faciais, como também, conseguimos identificar o interlocutor pela sua voz, mesmo quando alterada por problemas físicos ou emocionais.
- ▶ Combinar e empregar de maneira prática conhecimentos obtidos por meio de educação e experiências passadas. Por exemplo, um médico consegue diagnosticar um paciente combinando sintomas, resultados de exames clínicos e conhecimentos adquiridos durante sua formação e experiência profissional.

Introdução

- ▶ No início, quando a Inteligência Artificial (IA) começou a ser utilizada na solução de problemas reais, estes eram tratados pela IA por meio da aquisição de conhecimento de especialistas de um dado domínio (medicina, por exemplo) que era então codificado, frequentemente por regras lógicas, em um programa de computador. Esses programas eram conhecidos como Sistemas Especialistas ou Sistemas Baseados em Conhecimento.
- ▶ A aquisição do conhecimento de especialistas ocorria por meio de entrevistas que buscavam descobrir que regras eles utilizavam para tomar decisões. Esse processo possui várias limitações, como subjetividade, e, muitas vezes, pouca cooperação por parte do especialista, por causa do receio de ser dispensado após repassar seu conhecimento.
- ▶ A crescente complexidade dos problemas a serem computacionalmente tratados, e da velocidade e volume de dados gerados por diferentes setores, motivou o desenvolvimento de ferramentas computacionais mais sofisticadas e autônomas para a aquisição de conhecimento. A maioria dessas ferramentas é baseada em Aprendizado de Máquina (AM).

Introdução

- ▶ A Aprendizagem de Máquina (AM), também chamada de Machine Learning (ML), no inglês, nasceu na década de 60 como um campo da inteligência artificial;
- ▶ Inicialmente, as aplicações de AM eram de cunho estritamente computacional.
- ▶ Contudo, nos anos 90, essa área expandiu seus horizontes e começou a se estabelecer como um campo por si mesma.
- ▶ Em particular, as aplicações de AM começaram a ter muitas intersecções com as de estatística, com o uso de muitos modelos e métodos estatísticos para a solução dos problemas.



Aprendizagem de Máquina

O que é o Aprendizado de Máquina?

- ▶ Conjunto de regras e procedimentos (algoritmos, métodos, modelos.) que têm como objetivo identificar padrões (ou aprender) com os dados.

Para que serve?

- ▶ para produzir modelos capazes entregar resultados bastante precisos.



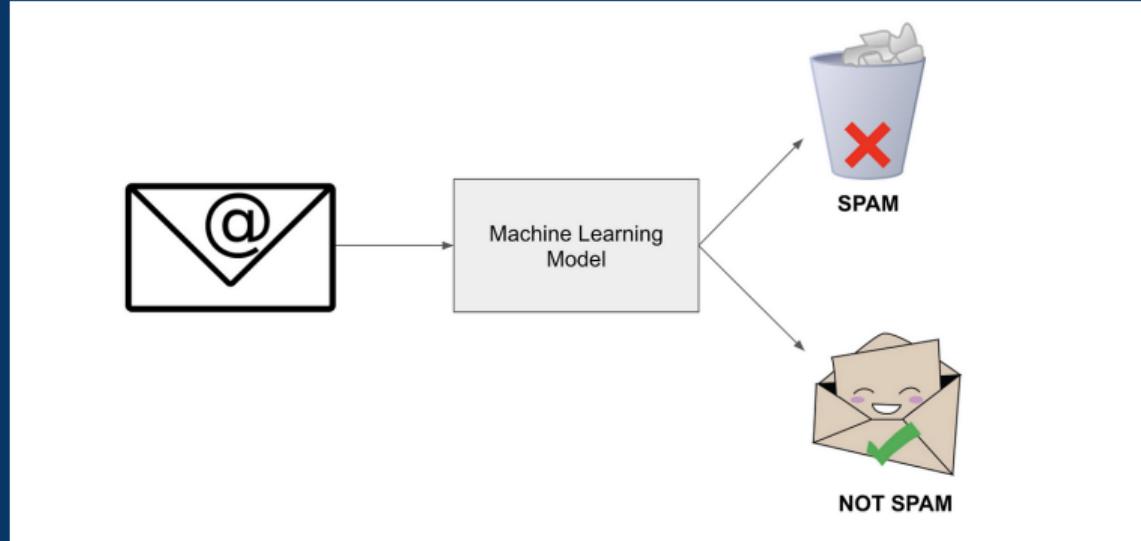
Fonte: <https://maquinasqueaprendem.com/2019/11/07/quando-usar-aprendizagem-de-maquina/>

Onde é usado?

- ▶ Na prática, podemos citar alguns exemplos reais:

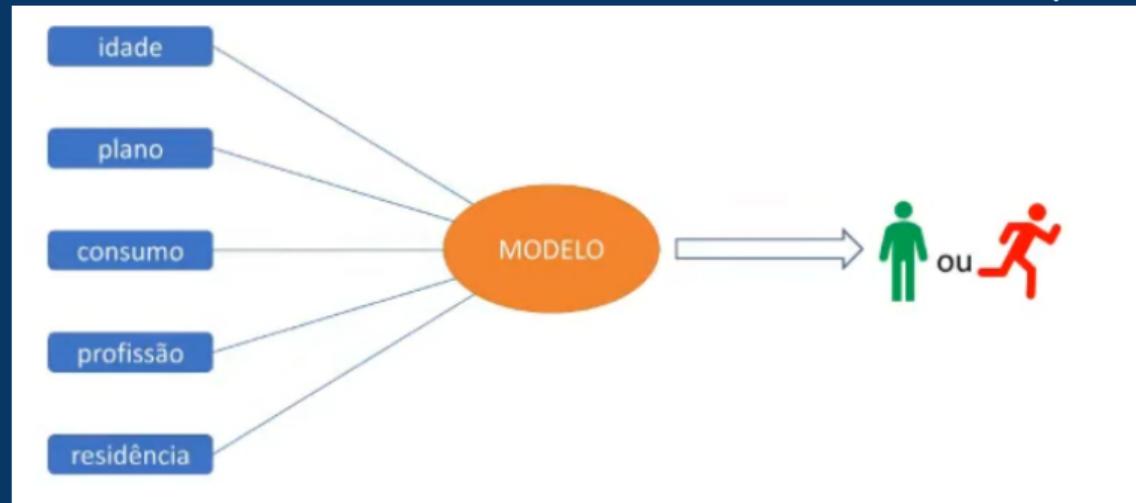


Detecção de Spam

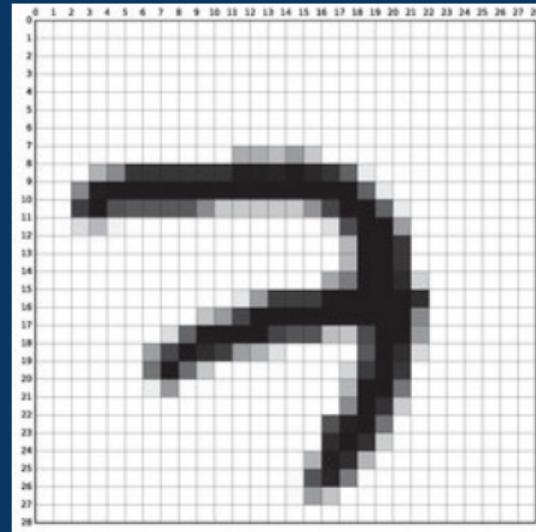


Cancelamento de um Serviço/Cliente

Prever a chance de um cliente deixar de consumir certo produto.



Reconhecimento de Dígito

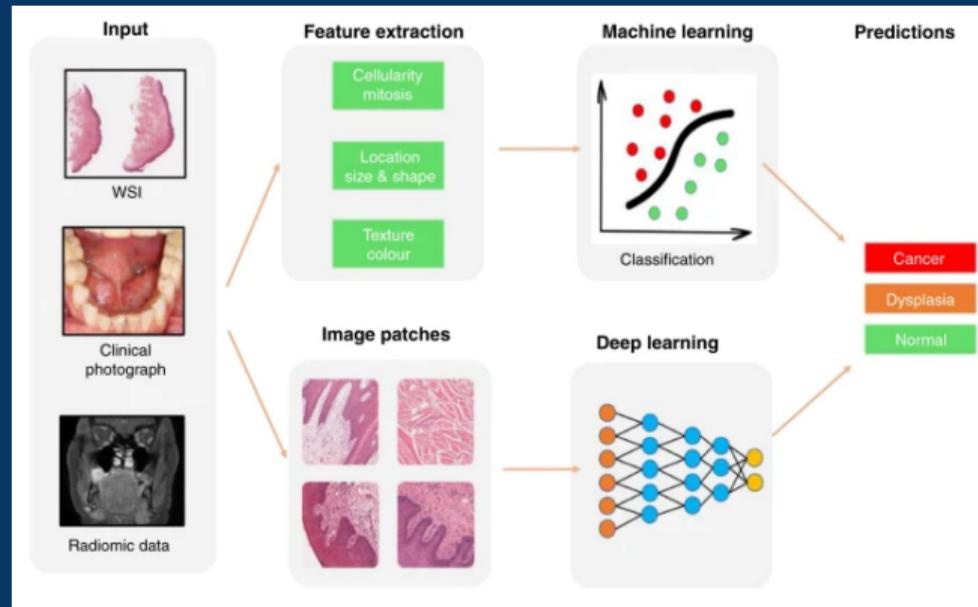


0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9



Diagnósticos

Por exemplo, quando há o interesse de reconhecer em uma imagem se há tecidos cancerígenos.



Fonte: <https://www.nature.com/articles/s41416-021-01386-x>

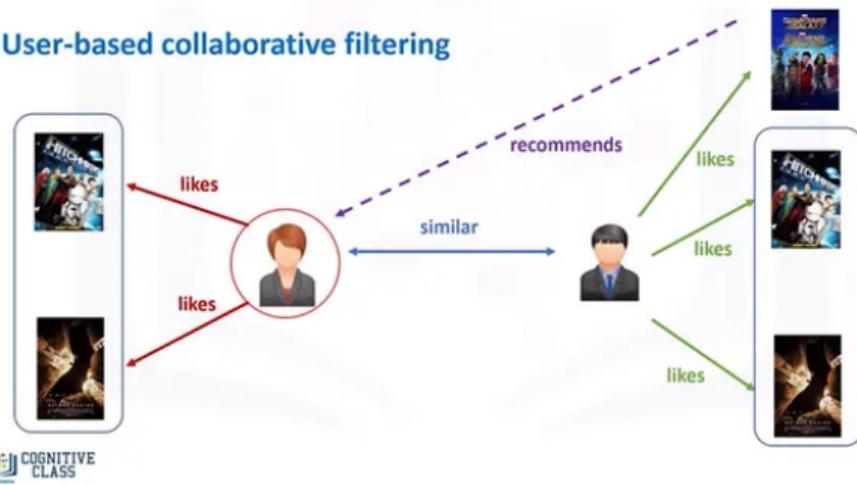


Sistemas de Recomendação

Filtro Colaborativo Baseado no Usuario

Collaborative filtering

- User-based collaborative filtering



Sistemas de Recomendação

Filtro Colaborativo Baseado no Produto

Content-based recommender systems



Tipos de Aprendizagem de Máquina

Existem diversas ferramentas em AM para a compreensão de dados. Vamos, inicialmente, dividir essas ferramentas em duas subáreas:

- ▶ **Aprendizagem supervisionada:** que consiste em aprender a fazer previsões a partir de conjunto de dados em que rótulos (ou seja, valores da variável resposta Y) são observados.
- ▶ **Aprendizagem não supervisionada:** que consiste em aprender as relações e a estrutura dos dados na ausência de rótulos (sem uma variável resposta Y).

Essas abordagens não são únicas e exclusivas, existem outros tipos de aprendizagem, como a semi-supervisionada e a por esforço, mas nosso foco será nas abordagens de aprendizagem supervisionada e não supervisionada.



O que é Aprender?

- ▶ De forma simples, aprender é ganhar conhecimento através do estudo, experiência ou sendo ensinado
- ▶ **Como a máquina aprende?**
 - ▶ **Aprendizagem** é o processo pelo qual se adquire o conhecimento → Algoritmos.
Ou seja, é o processo em que são utilizados algoritmos nos dados para extrair o conhecimento.
 - ▶ **Aprendizado** é o conhecimento adquirido → Modelos.
Ou seja, o modelo ajustado, obtido no processo de aprendizagem.
- ▶ Na disciplina de Aprendizagem de Máquina, focamos no estudo de **algoritmos** para adquirir descrições estruturais (**modelos**) sobre exemplos de dados



Aprendizagem Supervisionada

- ▶ O objetivo dos métodos de aprendizado supervisionado é, dadas as medições (instâncias/observações) $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$, aprender a predizer y_i baseado em \mathbf{x}_i .
- ▶ Ou seja, queremos um modelo para fazer previsões para uma variável de saída (que sabemos qual é) com base em uma ou mais variáveis de entrada.
- ▶ Em aprendizagem supervisionada os dados possuem um rótulo alvo conhecido (que é a variável resposta ou variável de saída, em inglês, *label*).
- ▶ Cada instância é então descrita por um vetor de atributos (medidas das variáveis preditoras, ou *features* em inglês), $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, e pelo rótulo (valor da variável resposta) associado y_i .
- ▶ As técnicas são aplicadas aos dados para ajustar (treinar, na linguagem AM) um modelo para predizer com precisão a resposta para novas observações com base nos atributos.
- ▶ Os modelos de regressão são exemplos de aprendizado supervisionado.



Aprendizagem não supervisionada

- ▶ Em aprendizagem não supervisionada o objetivo é mais genérico e, em geral, não tão bem especificado. Visto que existem variáveis de entrada, mas não há uma saída supervisionada.
- ▶ Dadas medições (instâncias/observações) $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, os métodos de aprendizado não supervisionado tentam descobrir alguma estrutura com base em similaridade, para determinar padrões previamente desconhecidos nos dados.
- ▶ Aqui, os dados não possuem rótulos (variável resposta) para ajudar no ajuste dos modelos. As instâncias são descritas pelos atributos (medidas das variáveis de entrada, ou *features* em inglês), $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$, $i = 1, \dots, n$.
- ▶ Os métodos de agrupamento (cluster), são exemplos de métodos de aprendizado não supervisionado.



Tarefas de Aprendizagem Supervisionada

- ▶ Os algoritmos de aprendizagem supervisionada são utilizados em tarefas preditivas.
- ▶ Em tarefas preditivas, algoritmos de AM são aplicados a conjuntos de dados de treinamento rotulados para produzir um modelo preditivo capaz de prever, para um novo objeto representado pelos valores de seus atributos preditivos, o valor de seu atributo alvo.
- ▶ Modelos preditivos podem ser utilizados, por exemplo, para, a partir dos sintomas de um paciente, prever o seu estado de saúde.
- ▶ O termo supervisionado vem da simulação da presença de um “supervisor externo”, que conhece, por exemplo, o verdadeiro diagnóstico do novo paciente. Essa informação é usada para guiar o processo de aprendizado na extração de um modelo com boa capacidade preditiva.

Tarefas de Aprendizagem Supervisionada

- ▶ As tarefas preditivas se distinguem pelo tipo do rótulo a ser predito: **Variável Quantitativa** ou **Variável Qualitativa (categorizada)**.
- ▶ As variáveis quantitativas assumem valores numéricos. Os exemplos incluem a idade, altura ou renda de uma pessoa, o valor de uma casa e o preço de uma ação.
- ▶ As variáveis qualitativas (categorizadas) assumem valores em uma das K classes ou categorias diferentes. Os exemplos incluem o estado civil (casado ou não), a marca do produto (marca A, B ou C), se uma pessoa não paga uma dívida (sim ou não) ou um diagnóstico de câncer (Leucemia Mielógena Aguda, Leucemia Linfoblástica Aguda ou Sem Leucemia).
- ▶ É comum se referir a problemas com resposta quantitativas como problemas de regressão, enquanto aqueles que envolvem uma resposta categorizada são frequentemente chamados de problemas de classificação.



Tarefas de Aprendizagem Supervisionada

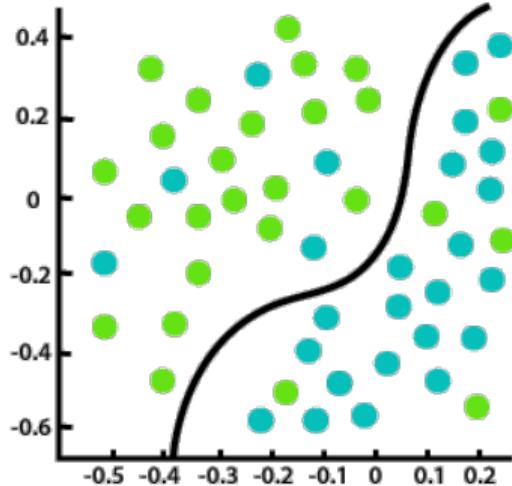
- ▶ Tipos de tarefas preditivas:
 - ▶ Classificação: determinar a classe de uma instância dados os seus valores atributos (*features*), i.e. $\hat{y} = \arg \max_y P(Y = y|X = \mathbf{x})$
 - ▶ Regressão: estimar o valor esperado da variável alvo de uma instância dados os seus atributos, i.e. $\hat{y} = \mathbb{E}[Y|X = \mathbf{x}]$
- ▶ No entanto, há casos que esta distinção nem sempre é tão clara. A regressão linear de mínimos quadrados é usada com uma resposta quantitativa, enquanto a regressão logística é normalmente usada com uma resposta qualitativa (duas classes ou binária). Assim, apesar do nome, a regressão logística, em AM, é um método de classificação (estima as probabilidades de classe).



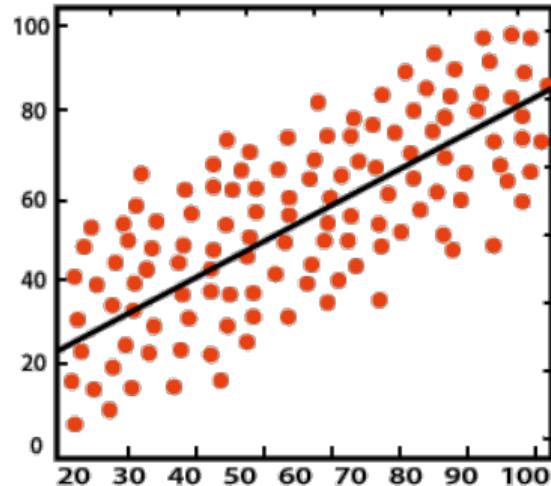
Tarefas de Aprendizagem Supervisionada

Os algoritmos de classificação podem ser divididos ainda em:

- ▶ **Generativos:** dadas as variáveis X e Y , o objetivo é encontrar a distribuição de probabilidade conjunta $P(X, Y)$ para a partir daí determinar $P(Y|X = \mathbf{x})$. Alguns métodos são:
 - ▶ Naive Bayes
 - ▶ Discriminante linear
- ▶ **Discriminativos:** buscam estimar diretamente a probabilidade condicional $P(Y|X = \mathbf{x})$ ou nem assumem modelos probabilísticos. Os modelos dessa classe são projetados para aprender a fronteira de decisão que separa as classes diretamente com base nas características de entrada. Podemos citar:
 - ▶ Regressão logística
 - ▶ Perceptron
 - ▶ Support Vector Machine- SVM



Classification

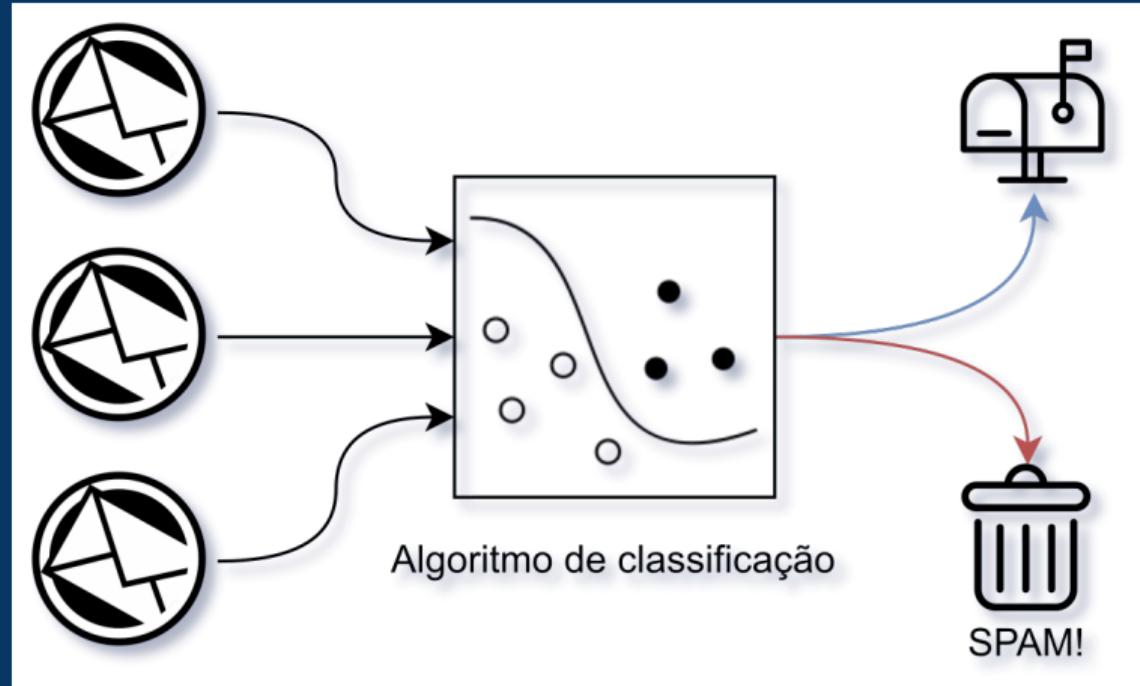


Regression

Alguns métodos estatísticos, como K-vizinhos mais próximos (KNN) e *boosting*, podem ser usados no caso de respostas quantitativas ou qualitativas.

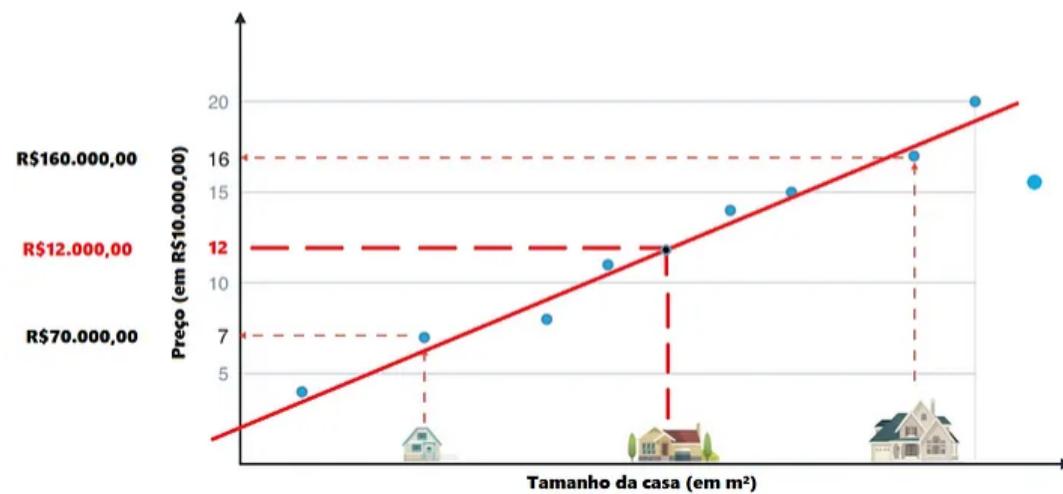


Exemplo: Classificação



Exemplo: Regressão

Preço de uma casa



Fonte: https://medium.com/@patrick_ams/regress%C3%A3o-linear-o-primeiro-passo-no-machine-learning-8f1c301f4529



Exemplo: Classificação

Tamanho (P)	Largura (P)	Tamanho (S)	Largura (S)	Especie
5,1	3,5	1,4	0,2	<i>Setosa</i>
4,9	3,0	1,4	0,2	<i>Setosa</i>
7,0	3,2	4,7	1,4	<i>Versicolor</i>
6,4	3,2	4,5	1,5	<i>Versicolor</i>
6,3	3,3	6,0	2,5	<i>Virginica</i>
5,8	2,7	5,1	1,9	<i>Virginica</i>



Exemplo: Regressão

No problema de classificação
CLASSE

↑

↓

↓

Objeto ou Observação →					
	Fertilidade	Agricultura	Educação	Renda	Mortalidad
	80,2	17,0	12	9,9	22,2
	83,1	45,1	9	84,8	22,2
	92,5	39,7	5	93,4	20,2
	85,8	36,5	7	33,7	20,3
	76,9	43,5	15	5,2	20,6

↓

Atributos preditivos,
Variáveis independentes,

↓

Atributo alvo,
Variável dependente,
Variável objetivo

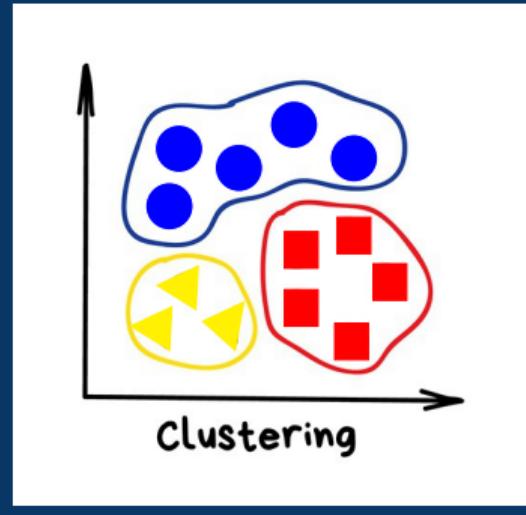
Tarefas de Aprendizagem não-Supervisionada

- ▶ Os algoritmos de aprendizagem não supervisionada são utilizados em tarefas chamadas descritivas.
- ▶ Em tarefas de descrição, ao invés de predizer um valor, algoritmos extraem padrões dos atributos preditivos de um conjunto de dados.
- ▶ Como não fazem uso do conhecimento do “supervisor externo”, esses algoritmos usam o paradigma de aprendizado não supervisionado.
- ▶ Uma das principais tarefas descritivas, agrupamento de dados, procura grupos de objetos similares entre si no conjunto de dados.
- ▶ Outra tarefa descritiva é encontrar regras de associação, que associam valores de um subconjunto de atributos preditivos a valores de outro subconjunto.



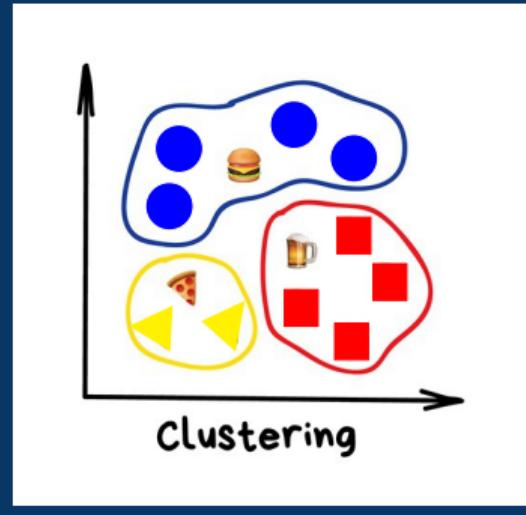
Agrupamento

- ▶ O algoritmo verifica se as instâncias observadas podem ser arranjadas de alguma maneira, formando grupos (*clusters*)
- ▶ O objetivo é que os clusters sejam maximamente parecidos internamente e maximamente diferentes entre si
- ▶ Após a determinação dos grupos, é necessário analisá-los para entender o que cada um representa



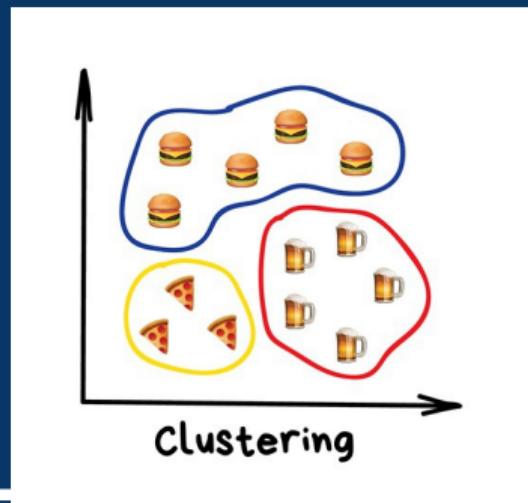
Agrupamento

- ▶ O algoritmo verifica se as instâncias observadas podem ser arranjadas de alguma maneira, formando grupos (*clusters*)
- ▶ O objetivo é que os clusters sejam maximamente parecidos internamente e maximamente diferentes entre si
- ▶ Após a determinação dos grupos, é necessário analisá-los para entender o que cada um representa

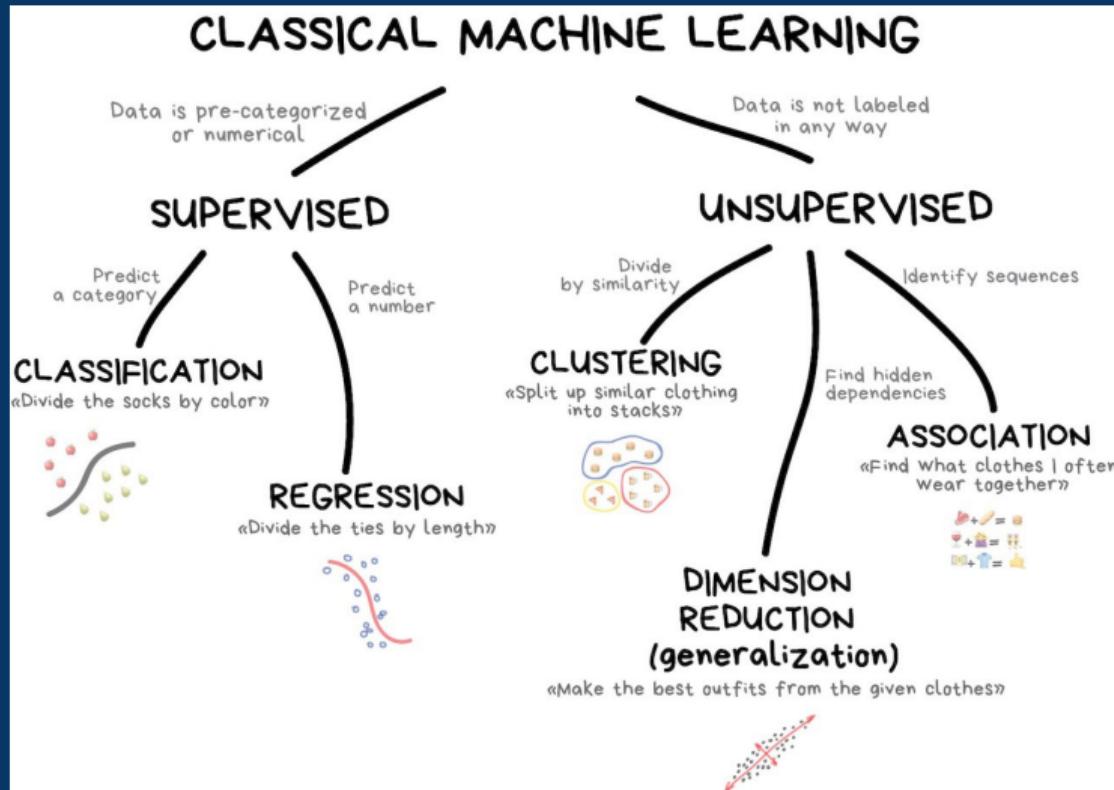


Agrupamento

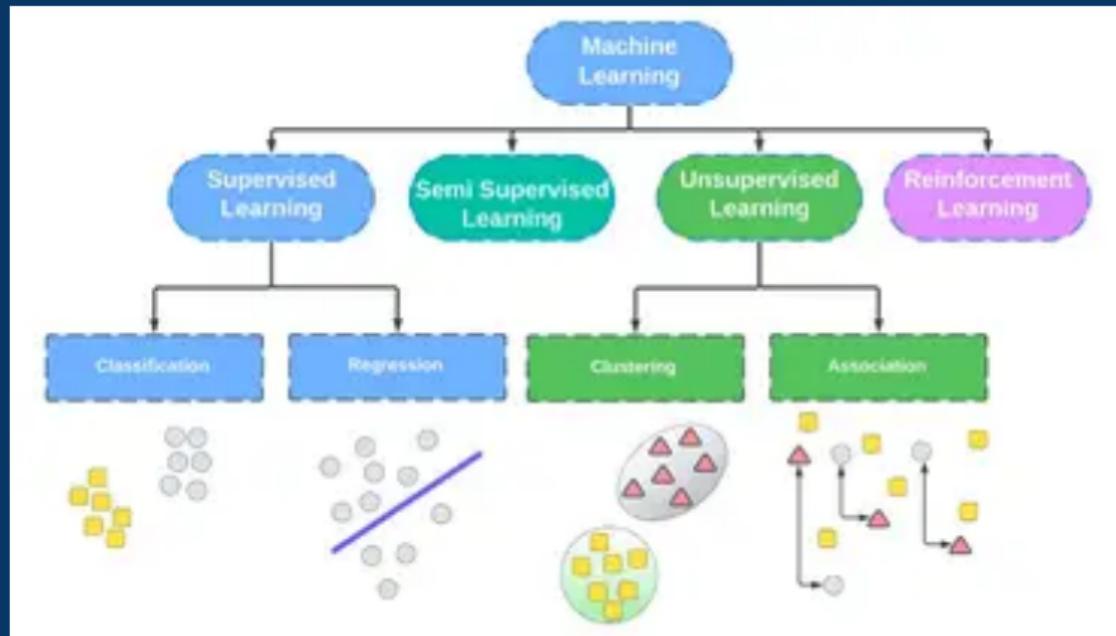
- ▶ O algoritmo verifica se as instâncias observadas podem ser arranjadas de alguma maneira, formando grupos (*clusters*)
- ▶ O objetivo é que os clusters sejam maximamente parecidos internamente e maximamente diferentes entre si (ou seja, precisam ser muito similares dentro dos grupos e muito dissimilares entre os grupos).
- ▶ Após a determinação dos grupos, é necessário analisá-los para entender o que cada um representa



Aprendizagem de Máquina



Aprendizagem de Máquina



Tipos de Aprendizagem de Máquina

- ▶ Aprendizagem supervisionada
- ▶ Aprendizagem não supervisionada
- ▶ Aprendizagem semi-supervisionada
- ▶ Aprendizagem por reforço



Aprendizagem Semi-Supervisionada

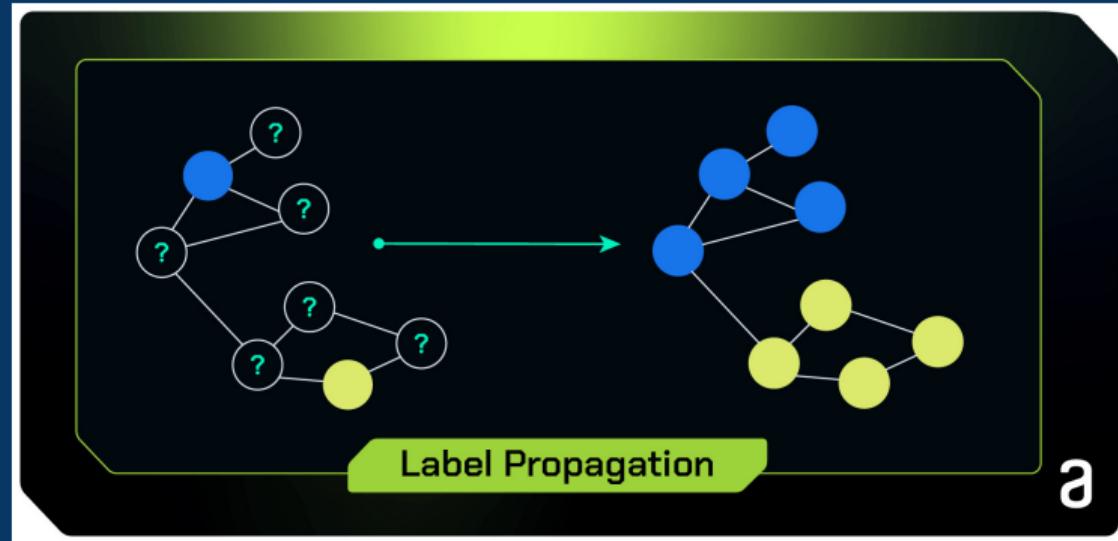
- ▶ Esse tipo de aprendizagem assume que o conjunto de treinamento possui instâncias rotuladas e (frequentemente muito mais) instâncias não-rotuladas
- ▶ O objetivo dos algoritmos semi-supervisionados é usar toda a informação possível,
- ▶ Em notação, usa-se as instâncias rotuladas para estimar $P(Y|X)$ e todas as instâncias, incluindo as não rotuladas, para estimar $P(X)$, tudo isso simultaneamente, de forma que uma estimativa ajude a outra
- ▶ Por exemplo, algoritmos que propagam rótulos, como o *Label Propagation*, em que rótulos conhecidos são propagados para dados não rotulados com base em sua proximidade no espaço de características.



Exemplo

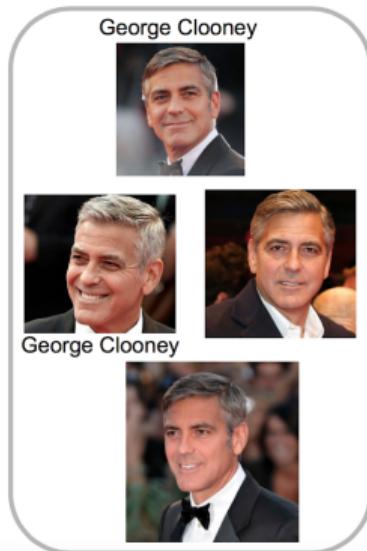
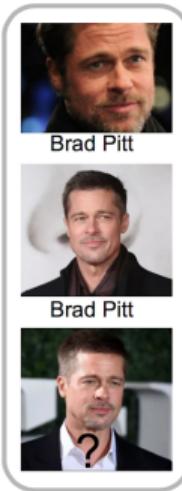


Exemplo



Exemplo

Exemplo

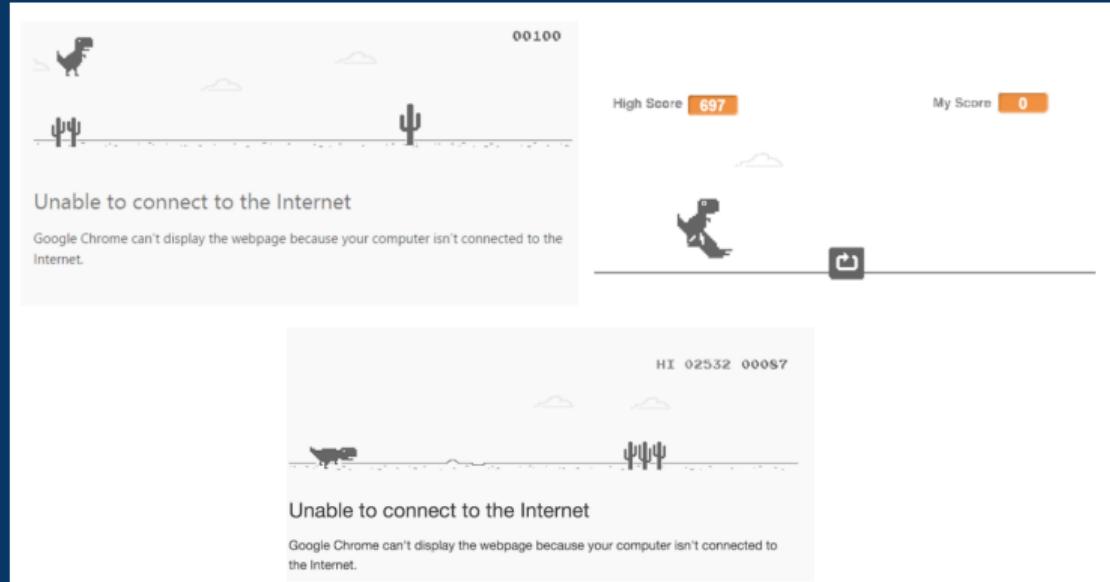


Aprendizagem por Reforço

- ▶ Aprendizagem por reforço (Sutton, R.S. e Barto, A.G., 1998) envolvem situações em que um ou mais agentes aprendem por tentativa e erro ao atuar sobre um ambiente dinâmico
- ▶ Não há uma fonte externa de exemplos. Há apenas a própria experiência do agente
- ▶ É necessário definir que ações o agente pode desempenhar e qual é a medida de desempenho

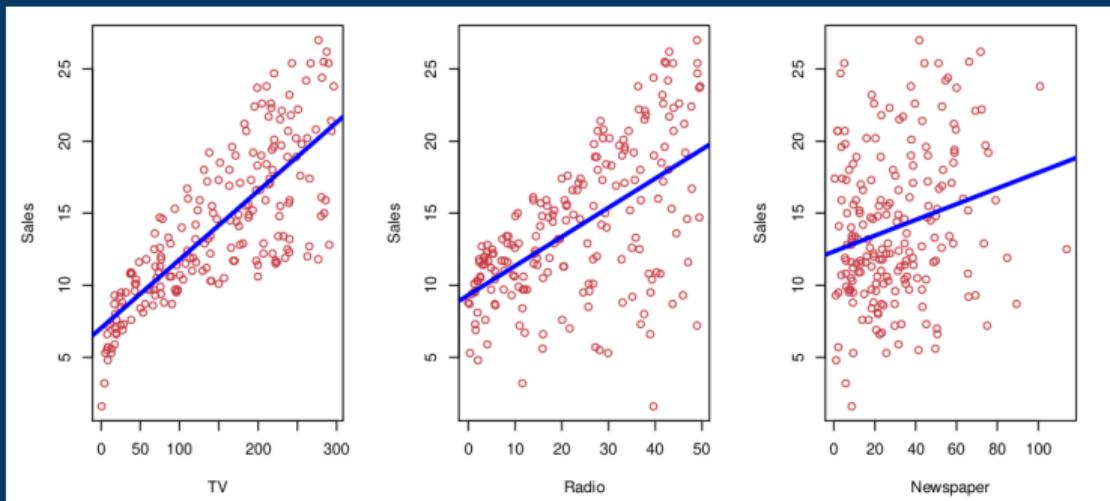


Exemplo



Motivando nosso estudo

- ▶ Suponha que fomos contratados por um cliente para investigar a associação entre publicidade e vendas de um determinado produto.
- ▶ O conjunto de dados consiste nas vendas (**Sales**) desse produto em diferentes mercados, juntamente com seus orçamentos de publicidade em três meios de comunicação diferentes: TV (**TV**), rádio (**Radio**) e jornal (**Newspaper**) .



Motivando nosso estudo

- ▶ Se determinarmos que existe uma associação entre publicidade e vendas, poderemos instruir nosso cliente a ajustar os orçamentos de publicidade, aumentando assim indiretamente as vendas.
- ▶ Em outras palavras, nosso objetivo é ajustar um modelo preciso que possa ser usado para prever vendas com base nos três orçamentos de mídia.

$$\text{Vendas} \approx f(\text{TV}, \text{Radio}, \text{jornal})$$

- ▶ Aqui, os orçamentos de publicidade são variáveis de entrada ($X_1 = \text{TV}$, $X_2 = \text{Radio}$, $X_3 = \text{jornal}$, também chamadas de variáveis preditoras, atributos ou *features*), enquanto as vendas é variável de saída ($Y = \text{Vendas}$, também chamada de variável resposta Y), a qual queremos prever.



Motivando nosso estudo

- De forma mais geral, suponha que observamos uma resposta quantitativa Y e p diferentes preditores, X_1, X_2, \dots, X_p . Assumimos que existe alguma relação entre Y e $\mathbf{X} = (X_1, X_2, \dots, X_p)$, que pode ser escrita na forma geral

$$Y = f(\mathbf{X}) + \epsilon$$

- Aqui f é alguma função fixa, mas desconhecida, de X_1, X_2, \dots, X_p , e ϵ é um termo de erro aleatório que tem média zero.
- Nesta fórmula, f representa a informação sistemática que \mathbf{X} fornece sobre Y .



Motivando nosso estudo

- ▶ No entanto, a função f que conecta as variáveis de entrada à variável de saída é em geral desconhecida.
- ▶ Dessa forma, é preciso estimar f com base nos valores observados.
- ▶ Em essência, a aprendizagem de máquina refere-se a um conjunto de abordagens para estimar f .
- ▶ Descreveremos alguns dos principais conceitos teóricos para estimar f , bem como ferramentas para avaliar as estimativas obtidas.



Porque estimar f ?

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

Predição

- ▶ Podemos prever Y usando

$$\hat{Y} = \hat{f}(X)$$

- ▶ onde \hat{f} representa a estimativa para f e \hat{Y} representa a previsão para Y .
- ▶ Na predição, não estamos muito preocupados com a forma exata de \hat{f} , desde que produza previsões precisas para Y .
- ▶ Ou seja, nesta configuração, \hat{f} é frequentemente tratado como uma "*caixa preta*", no sentido de que não há a preocupação com a forma exata de \hat{f} , desde que produza previsões precisas para Y .



Inferência

- ▶ Nesta situação desejamos estimar f mais interessados em compreender a associação entre Y e X_1, X_2, \dots, X_p , do que necessariamente fazer previsões para Y .
- ▶ Agora \hat{f} não pode ser tratada como uma "*caixa preta*", porque precisamos saber a sua forma exata.
- ▶ Neste cenário, podemos estar interessado em:
 - ▶ Identificar quais preditores estão associados a resposta.
 - ▶ Qual a forma estabelecida entre Y e os preditores.
 - ▶ Como cada preditor se relaciona com a resposta.



Motivando nosso estudo

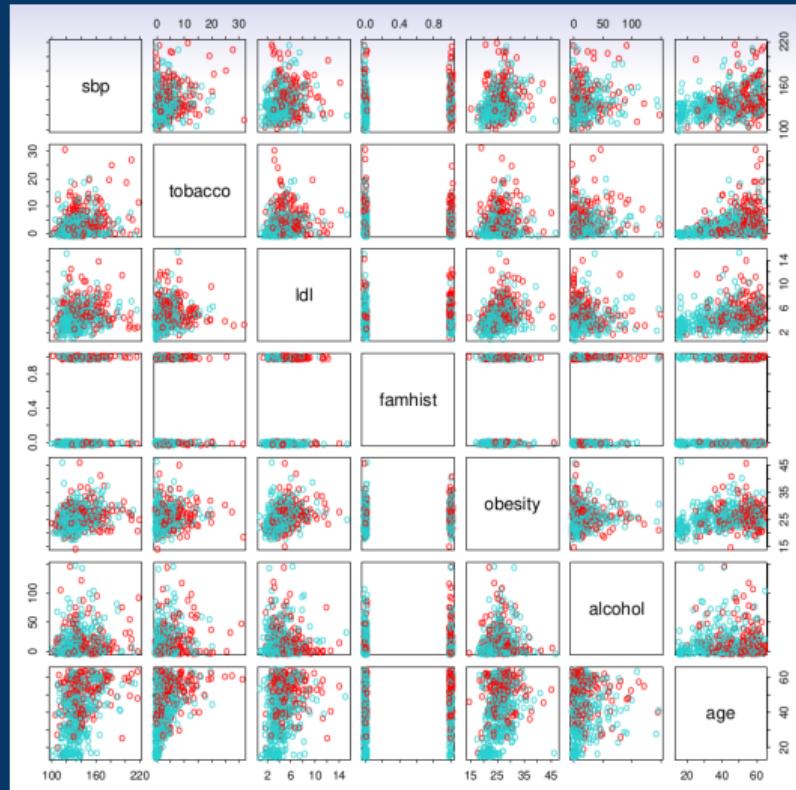
Dois Objetivos

- ▶ Podemos dividir as razões para estimar um modelo em dois principais objetivos
 - ▶ **Objetivo inferencial:** Quais preditores são importantes? Qual a relação entre cada preditor e a variável resposta? Qual o efeito da mudança de valor de um dos preditores na variável resposta?
 - ▶ **Objetivo preditivo:** Como podemos criar uma função que tenha bom poder preditivo?

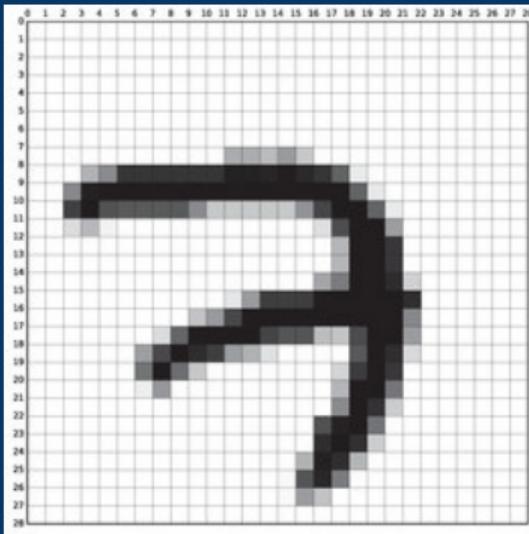
Enquanto em alguns dos problemas o objetivo é claramente inferencial ou preditivo, em outros pode uma mistura de ambos.



Objetivo Inferencial: Identificar fatores de risco para um ataque cardíaco



Objetivo preditivo: Reconhecimento de dígito



0 0 0 0 0 0 0 0 0
1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3
4 4 4 4 4 4 4 4 4
5 5 5 5 5 5 5 5 5
6 6 6 6 6 6 6 6 6
7 7 7 7 7 7 7 7 7
8 8 8 8 8 8 8 8 8
9 9 9 9 9 9 9 9 9



Motivando nosso estudo

- ▶ Dependendo se o nosso objetivo final é a previsão, a inferência ou uma combinação dos dois, diferentes métodos para estimar f podem ser apropriados.
- ▶ Por exemplo, os modelos lineares permitem inferências relativamente simples e interpretáveis, mas podem não produzir previsões tão precisas como algumas outras abordagens.
- ▶ Em contraste, algumas das abordagens altamente não lineares (que discutiremos posteriormente) podem potencialmente fornecer previsões bastante precisas para Y , mas isto ocorre à custa de um modelo menos interpretável, para o qual a inferência é mais desafiadora.



Como estimar f

- Existem muitas abordagens lineares e não lineares para estimar f . Estudaremos alguns desses métodos.
- Iniciaremos apresentando as notações e características comuns para os dois métodos.

Notações

- Temos um conjunto de n observações, que serão chamadas de dados de treinamento porque usaremos essas observações para treinar, ou ensinar, nosso método como estimar f .
- x_{ij} representará o valor do j -ésimo preditor, ou entrada, para a observação i , onde $i = 1, 2, \dots, n$ e $j = 1, 2, \dots, p$,
- y_i vai representar a variável de resposta para a i -ésima observação.
- Os dados de treinamento consistem em $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ onde $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$.

Como estimar f

- ▶ Nosso objetivo é aplicar um método de aprendizagem de máquina 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) .
- ▶ Os métodos para treinar, ou estimar f podem ser divididos entre **paramétricos** ou **não paramétricos**.



Métodos Paramétricos

Os métodos paramétricos envolvem uma modelagem baseada em duas etapas.

1. Fazemos uma suposição sobre a forma funcional de f . Uma suposição muito simples é que f é linear em X :

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

Uma vez assumido que f é linear, o problema de estimar f é o de estimar os coeficientes do modelo.

2. Após a seleção de um modelo, precisamos de um procedimento que use os dados de treinamento para ajustar ou treinar o modelo. Ou seja, treinar para estimar os parâmetros.

A abordagem mais comum para ajustar o modelo linear é por **mínimos quadrados**.



- ▶ A abordagem baseada em um modelo, que acabamos de descrever, é chamada de paramétrica e reduz o problema de estimar f a estimar um conjunto de parâmetros.
- ▶ Assumir uma forma paramétrica para f simplifica o problema de estimar f , porque geralmente é muito mais fácil estimar um conjunto de parâmetros, como $\beta_0, \beta_1 \dots, \beta_p$ no modelo linear, do que ajustar uma função inteiramente arbitrária.
- ▶ A desvantagem de uma abordagem paramétrica é que o modelo que escolhemos geralmente não corresponderá à verdadeira forma desconhecida de f . E, se o modelo escolhido estiver longe do verdadeiro f , então nossa estimativa será ruim.
- ▶ É possível tentar resolver esse problema escolhendo modelos flexíveis que possam se ajustar a diferentes formas funcionais de f . Mas, em geral, ajustar um modelo mais flexível requer estimar um maior número de parâmetros.
- ▶ Esses modelos mais complexos podem levar a um fenômeno conhecido como *overfitting* dos dados (ou superajuste), o que significa essencialmente que eles seguem os erros ou ruídos muito de perto.

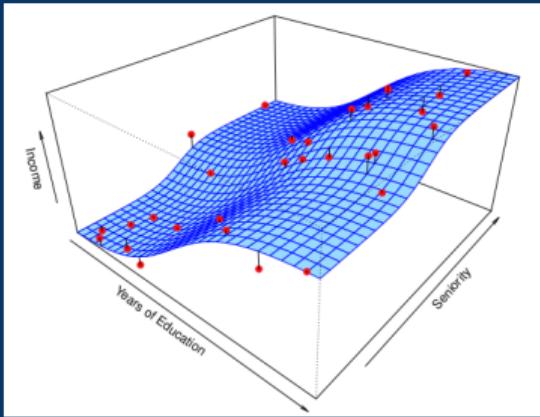


Métodos Não Paramétricos

- ▶ Os métodos não paramétricos não fazem suposições explícitas 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.
- ▶ Ao evitar a suposição de uma forma funcional específica para f , as abordagens não paramétrica tornam-se vantajosas sobre as abordagens paramétricas.
- ▶ A abordagem paramétrica traz consigo a possibilidade de especificar uma forma funcional muito diferente da verdadeira f , o que pode resultar em um modelo que não se ajuste bem aos dados.
- ▶ As abordagens não paramétricas evitam completamente este perigo, uma vez que não é feita nenhuma suposição sobre a forma de f .
- ▶ Mas 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, é necessário um número muito grande de observações (muito mais do que é normalmente necessário para uma abordagem paramétrica) para obter uma estimativa precisa para f .



Exemplo Simulado: Verdadeira Relação



- ▶ Os pontos vermelhos são valores simulados de renda do modelo

$$\text{Income} = f(\text{years of education}, \text{seniority}) + \epsilon$$

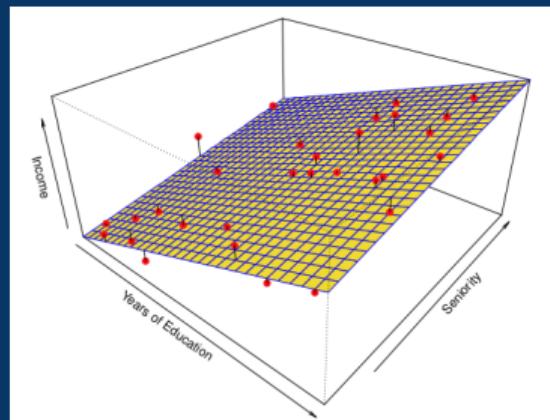
- ▶ f é a superfície azul, conhecida, já que são de dados simulados.



Exemplo Simulado: Modelo linear

- ▶ Um modelo linear ajustado por mínimos quadrados.

$$\hat{f}(\text{education}, \text{seniority}) = \beta_0 + \beta_1 \text{education} + \beta_2 \text{seniority}$$

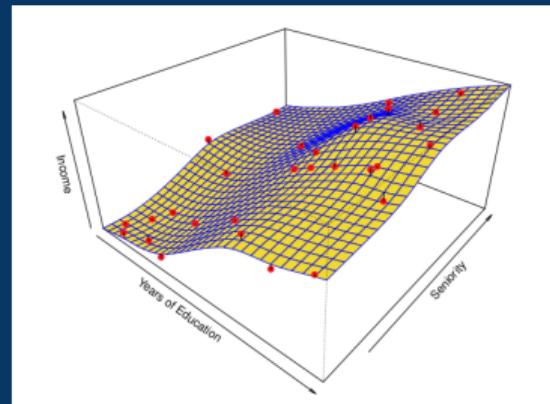


- ▶ As observações são mostradas em vermelho e o plano amarelo indica o modelo ajustado aos dados por mínimos quadrados.



Exemplo Simulado: Uma abordagem não paramétrica

- ▶ Uma abordagem não paramétrica utilizada para ajustar os dados (*thin-plate spline*).
- ▶ Esta abordagem não impõe nenhum modelo pré-especificado em f . Em vez disso, tenta produzir uma estimativa para f que seja o mais próxima possível dos dados observados

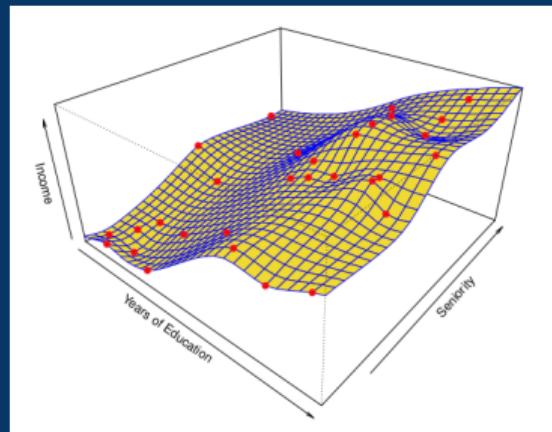


- ▶ $\hat{f}_{\text{flex}}(\text{education}, \text{seniority})$
- ▶ As observações são mostradas em vermelho e a superfície amarela indica o modelo mais flexível ajustado aos dados.
- ▶ Neste caso, o ajuste não paramétrico produziu uma estimativa notavelmente mais precisa do verdadeiro f .



Exemplo Simulado: *overfitting*

- Modelo não paramétrico ainda mais flexível ajustado aos dados simulados.
- A estimativa resultante ajusta-se perfeitamente aos dados observados!



- $\hat{f}_S(\text{education}, \text{seniority})$
- Aqui o modelo ajustado não comete erros nos dados de treinamento!
- O modelo ajustado fica muito mais variável do que a verdadeira função f .
- Este é um exemplo de *overfitting* dos dados.
- É uma situação indesejável porque o ajuste obtido não produzirá estimativas precisas da resposta em novas observações que não faziam parte do conjunto de dados de treinamento original.

Balanço entre a Flexibilidade e a Interpretabilidade do Modelo

- ▶ Dos muitos métodos que estudaremos, alguns são menos flexíveis, ou mais restritivos, no sentido de que podem produzir poucas formas para estimar f .
- ▶ Por exemplo, a regressão linear é uma abordagem relativamente inflexível (restrita), porque só pode gerar funções lineares.
- ▶ Outros métodos, como *thin-plate spline* mostrados no exemplo anterior, são consideravelmente mais flexíveis porque podem gerar um muito mais formas para estimar f .
- ▶ Poderíamos razoavelmente fazer a seguinte pergunta: por que escolheríamos usar um método mais restritivo em vez de uma abordagem muito flexível?



- ▶ Existem vários motivos pelos quais podemos preferir um modelo mais restritivo. Por exemplo, se estivermos interessados principalmente em inferência, então os modelos restritivos são muito mais interpretáveis.
- ▶ Quando o objetivo é a inferência, o modelo linear pode ser uma boa escolha, pois será bastante fácil entender a relação entre Y e X_1, X_2, \dots, X_p .
- ▶ Em contraste, abordagens muito flexíveis, podem levar a estimativas de f tão complicadas que é difícil entender como qualquer preditor individual está associado à resposta.
- ▶ Estabelecemos que quando o objetivo é a inferência, há vantagens em usar métodos simples e relativamente inflexíveis. Contudo, quando estamos interessados apenas na predição, e a interpretabilidade não é uma preocupação, neste cenário, é melhor usar o modelo mais flexível.



Balanço entre a Flexibilidade e a Interpretabilidade do Modelo



Duas Culturas

Breiman (2001a) argumenta que existem duas culturas no uso de modelos estatísticos

- ▶ ***Data modeling culture:*** é a cultura que domina a comunidade estatística. Isso ocorre pois o principal objetivo está na interpretação dos parâmetros envolvidos no modelo; em particular há interesse em testes de hipóteses e intervalos de confiança para esses parâmetros. Sob essa abordagem, testar se as suposições do modelo (por exemplo, normalidade dos erros, linearidade, homocedasticidade etc) são válidas é de fundamental importância. Ainda que predição muitas vezes faça parte dos objetivos, o foco em geral está na inferência.
- ▶ ***algorithmic modeling culture:*** é a que domina a comunidade de aprendizado de máquina (*machine learning*). Neste meio, o principal objetivo é a predição de novas observações. Não se assume que o modelo utilizado para os dados é correto; o modelo é utilizado apenas para criar bons algoritmos para prever bem novas observações. Muitas vezes não há nenhum modelo probabilístico explícito por trás dos algoritmos utilizados.

Duas Culturas

- ▶ Mesmo que o objetivo primordial de uma problema de predição seja obter um bom poder preditivo, a interpretabilidade do modelo final também é importante por diversas razões. Discutiremos mais adiante algumas dessas razões.
- ▶ Apesar dessa divisão de culturas, Breiman foi um estatístico que fez um grande trabalho para unir a área de estatística com aprendizado de máquina. Devido à sua grande importância nessa tarefa, um prêmio concedido em sua homenagem foi criado pela *American Statistical Association*.
- ▶ Essa união entre as áreas é mutuamente benéfica e, portanto, seguiremos tentando unir as duas culturas.



Avaliando a precisão (acurácia) do modelo

- ▶ Neste curso estudaremos diversos métodos de aprendizagem de máquina que vão muito além da abordagem tradicional de regressão linear.
- ▶ É muito importante saber diferentes tipos de abordagem, pois não há apenas um único método melhor, que domine todos os outros.
- ▶ Num conjunto de dados específico, um método específico pode funcionar melhor, mas algum outro método pode funcionar melhor em outro conjunto de dados, mesmo que estes dados sejam similares.
- ▶ Portanto, é uma tarefa importante decidir, para qual conjunto de dados, qual método produz os melhores resultados.
- ▶ Dessa forma, avaliar os modelos para selecionar a melhor abordagem pode ser uma das partes mais desafiadoras da realização do aprendizado de máquina na prática.



Medindo a qualidade do ajuste

- ▶ Para avaliar o desempenho de um método de aprendizagem de máquina em um determinado conjunto de dados, precisamos de alguma forma de medir até que ponto suas previsões realmente correspondem aos dados observados.
- ▶ Ou seja, precisamos quantificar até que ponto o valor da resposta prevista para uma determinada observação está próximo do valor verdadeiro da resposta para essa observação.
- ▶ Em regressão, a medida mais comumente usada é o erro quadrático médio (*MSE*), dado por:

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

- ▶ Um valor pequeno para o *MSE* indica que as respostas previstas estão muito próximas das respostas verdadeiras.

- ▶ Como o *MSE* é calculado usando os dados de treinamento (que foram usados para ajustar o modelo), ele pode ser uma medida "otimista"(podemos chamá-lo de *MSE de treinamento*).
- ▶ No geral, não é tão importante com o quanto bem o método funciona nos dados de treinamento, em vez disso, estaremos interessados na precisão das previsões quando aplicamos o nosso método a dados nunca antes vistos.
- ▶ Considere o caso que temos medidas clínicas (por exemplo, peso, pressão arterial, altura, idade, histórico familiar de doença) para vários pacientes, bem como informações sobre se cada paciente tem diabetes. Podemos usar esses pacientes para treinar um método de aprendizagem estatística para prever o risco de diabetes com base em medições clínicas.
- ▶ Mas na prática, queremos que este método preveja com precisão o risco de diabetes para futuros pacientes com base nas suas medições clínicas. Não estamos muito interessados em saber se o método prevê ou não com precisão o risco de diabetes para os pacientes usados para treinar o modelo, uma vez que já sabemos quais desses pacientes têm diabetes.



- Uma maneira de solucionar este problema é dividir o conjunto de dados em duas partes, **treinamento** e **validação**:

Treinamento (por exemplo, 70%) Validação (por exemplo, 30%)

$$\overbrace{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_s, Y_s)}^{\text{Treinamento}}, \quad \overbrace{(\mathbf{X}_{s+1}, Y_{s+1}), \dots, (\mathbf{X}_n, Y_n)}^{\text{Validação}}.$$

- Usamos o conjunto de treinamento exclusivamente para estimar f (por exemplo, estimar os coeficientes da regressão linear) e o conjunto de validação apenas para calcular

$$MSE_V = \frac{1}{n-s} \sum_{i=s+1}^n (y_i - \hat{f}(x_i))^2.$$

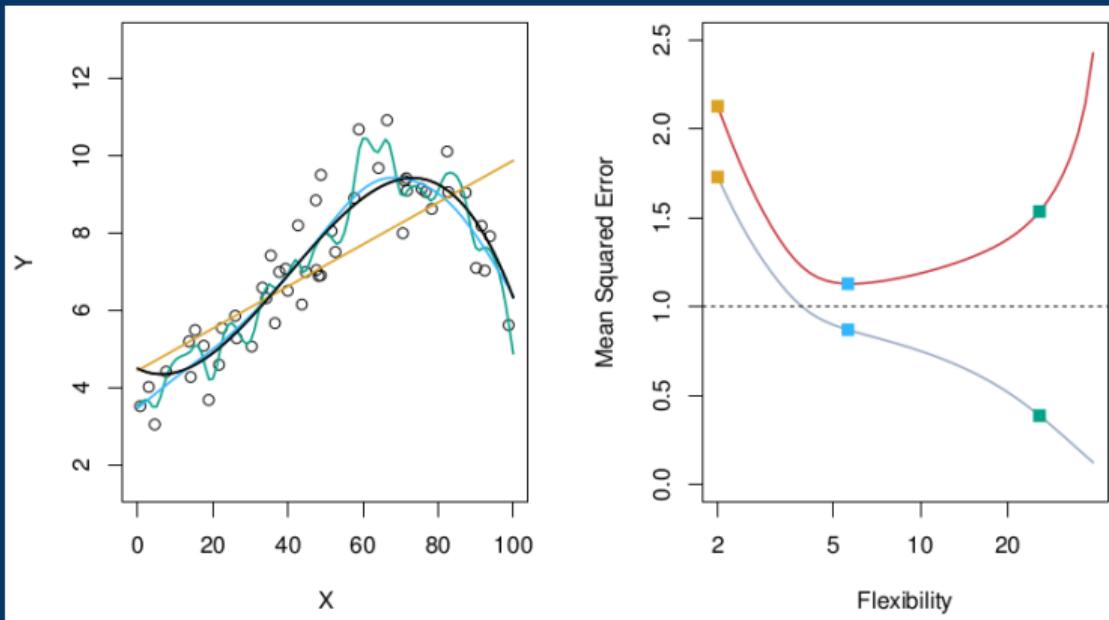
- isto é, avaliamos o erro quadrático médio no conjunto de validação.
- Este procedimento é chamado de *data splitting*. Existem outras variações deste método como a **validação cruzada**, que será visto mais adiante.



Seleção de Modelos: Super e Sub-Ajuste

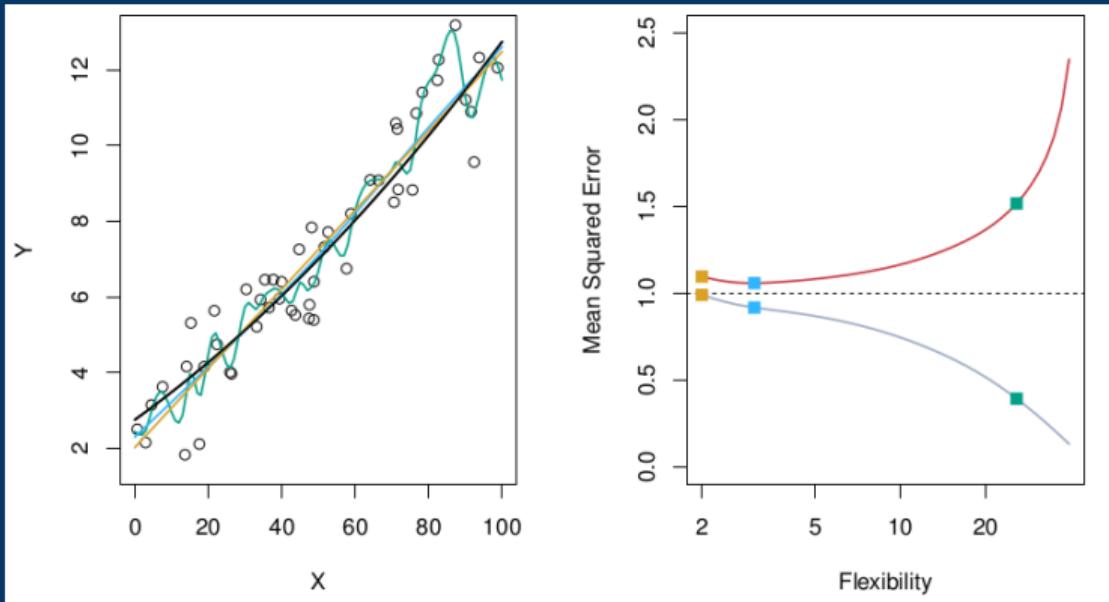
- ▶ O objetivo de um método de seleção de modelos é selecionar uma boa função f .
- ▶ Isso inclui evitar os seguintes problemas:
 - ▶ Super-ajuste (*Overfitting*): O modelo se ajusta demais a uma amostra específica, mas possui baixo poder de generalização.
 - ▶ Sub-ajuste (*Underfitting*): O modelo não é suficiente para explicar bem os dados.
- ▶ Para seleção de modelos, é comum ajustar vários modelos para a f e buscar qual deles possui maior poder preditivo.
- ▶ Utilizar o critério do erro quadrático médio no conjunto de validação, selecionando o método que minimiza essa quantidade (apresenta as melhores previsões para novas observações), pode evitar problemas de *Overfitting*.





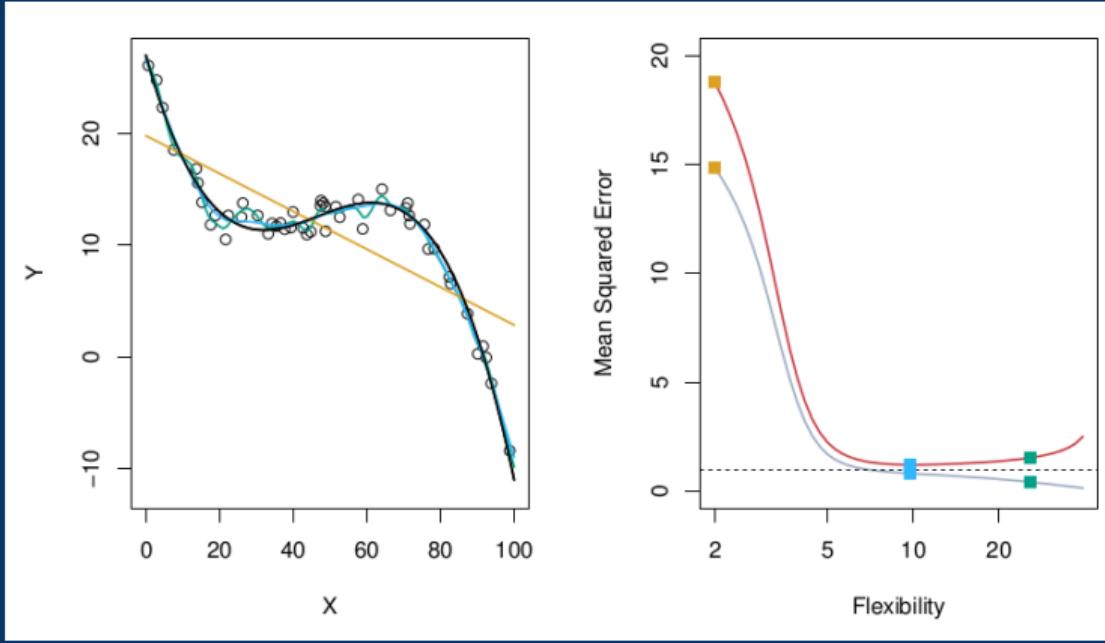
- ▶ Dados simulados de f em preto (verdadeira função).
- ▶ Três estimativas de f são mostradas (amarela, azul e verde).
- ▶ MSE de treinamento (curva cinza), MSE de teste (curva vermelha)





- ▶ Dados simulados com uma função verdadeira diferente para f (em preto), mas próxima da regressão linear.
- ▶ Nessa configuração, a regressão linear fornece um ajuste muito bom aos dados.





- Dados simulados com outra função verdadeira para f (em preto), mas distante da regressão linear.
- Neste cenário, a regressão linear fornece um ajuste muito fraco aos dados (Sub-ajuste ou Underfitting)



Balanço entre viés e variância

- ▶ Podemos decompor o *MSE* de teste, para um determinado valor \mathbf{x}_0 , como a soma de três quantidades fundamentais:

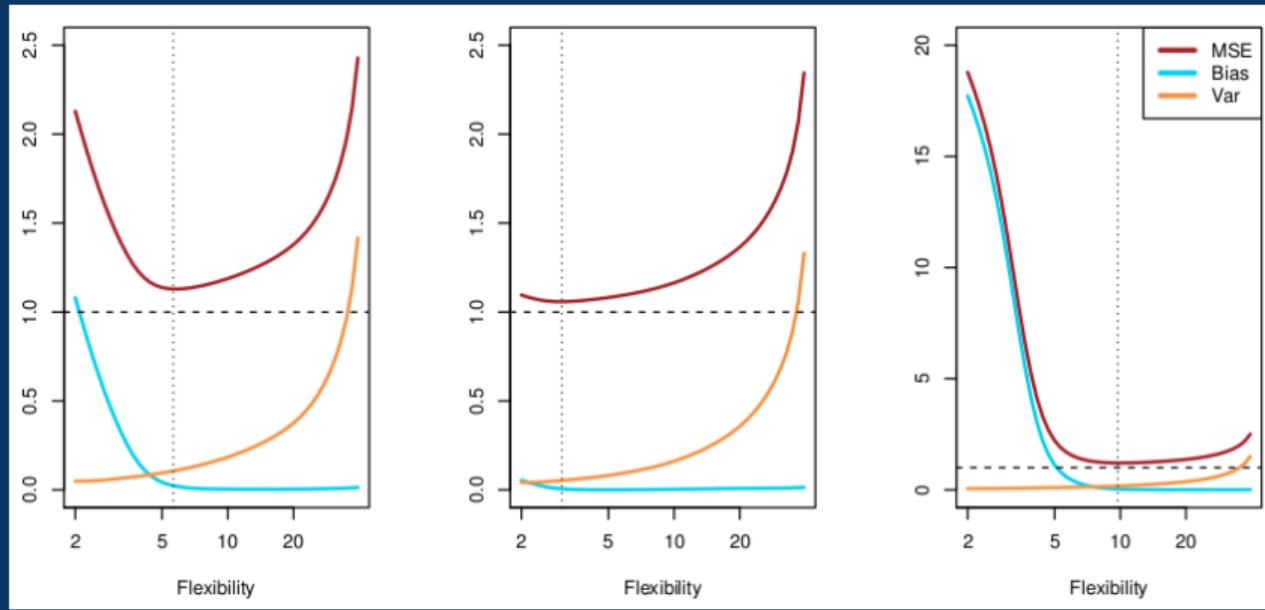
- ▶ a variância de $\hat{f}(\mathbf{x}_0)$,
- ▶ viés quadrático de $\hat{f}(\mathbf{x}_0)$
- ▶ a variância do erro

$$E \left(y_0 - \hat{f}(\mathbf{x}_0) \right)^2 = \text{Var} \left(\hat{f}(\mathbf{x}_0) \right) + \text{Bias} \left(\hat{f}(\mathbf{x}_0) \right)^2 + \text{Var}(\epsilon).$$

- ▶ Note que estamos trabalhando com o erro esperado sobre todos os valores possíveis de \mathbf{x}_0 no conjunto de teste.
- ▶ Para minimizar o erro esperado no conjunto de teste, precisamos selecionar um método de aprendizagem estatística que alcance simultaneamente baixa variância e baixo viés.
- ▶ Esse dois itens podem ser reduzidos se escolhermos \hat{f} adequado.



- ▶ Normalmente, à medida que a flexibilidade de \hat{f} aumenta, sua variância aumenta e seu viés diminui.
- ▶ Portanto, escolher a flexibilidade com base no erro quadrático médio de teste equivale a equilibrar esse balanço entre viés e variância.



Duas Culturas



Os Benditos Dados



Explicando os dados

- ▶ Atributos podem ser físicos ou abstratos, como sintomas
- ▶ Cada objeto/instância é descrito por um conjunto de atributos de entrada ou vetor de características
- ▶ Cada objeto corresponde a uma ocorrência/observação
- ▶ Os atributos estão associados a propriedades dos objetos

Conjuntos de dados

- ▶ Os dados que usamos para treinar nossos modelos são agregados em um **conjunto ou base de dados** (*data set* ou *dataset*)
- ▶ O conjunto de dados costuma ser representado por uma matriz $\mathbf{X}_{n \times p}$
 - ▶ n é o número de instâncias
 - ▶ p é o número de atributos de cada instância e define a dimensionalidade do espaço do problema



Exemplo

	Identificador	Nome	Idade	Sexo	Peso	Manchas	Temperatura	Internacoes	Estado	Diagnostico
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
5	2301	Ana	22	F	72	Inexistentes	38.0	3	RJ	Doente
6	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
7	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel



Conjuntos hospital

- ▶ Este conjunto aparentemente tem $d = 10$ variáveis
- ▶ No entanto, a primeira e a segunda são apenas identificadores de paciente
- ▶ E a última, que indica o diagnóstico, possivelmente será selecionada como alvo em uma tarefa de classificação, sendo tratada como uma variável separada Y
- ▶ Portanto, $d = 7$



Pré-processamento

- Antes de alimentar um algoritmo de aprendizagem de máquina com o conjunto de dados observados, comumente precisamos realizar diversas atividades de preparação dos dados, incluindo:
 - ▶ Eliminação manual de atributos
 - ▶ Integração de dados
 - ▶ Amostragem
 - ▶ Balanceamento
 - ▶ Limpeza
 - ▶ Redução de dimensionalidade
 - ▶ Transformação

Pré-processamento

- ▶ Essas técnicas são usadas para melhorar a qualidade dos dados, i.e. tornar mais fácil o ajuste de modelos
- ▶ Minimizam problemas de ruídos, anomalias/outliers, valores/rótulos incorretos, duplicados ou ausentes
- ▶ Também podem adequar os dados para uso de determinados algoritmos, e.g. algoritmos com entradas exclusivamente numéricas



Eliminação manual de atributos

- Removemos atributos que não contribuem para a construção dos modelos
- Nesse momento, o conhecimento e a experiência dos especialistas são fundamentais

Removendo colunas

	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico								
0	28	M	79	Concentradas	38.0	2	SP	Doente								
1	18	F	67	Inexistentes	39.5	4	MG	Saudavel								
2	49	M	92	Espalhadas	38.0	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico	
3	18	M	43	Inexistentes	38.5	0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente
4	21	F	52	Uniformes	37.6	1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
5	22	F	72	Inexistentes	58.0	2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
6	19	F	87	Espalhadas	39.0	3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
7	34	M	67	Uniformes	38.4	4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
						5	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
						6	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
						7	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel



Integração de dados

- ▶ Essa atividade trata da junção de duas bases de dados que possuem informações sobre os mesmos objetos
- ▶ Devemos buscar atributos comuns nos conjuntos que serão combinados
 - ▶ Exemplos: CPF, CNPJ e identificadores de uma maneira geral, além de outros atributos que podem estar repetidos
- ▶ Atributos cruzados devem ter um valor único para cada objeto

Amostragem de dados

- ▶ Alguns algoritmos de aprendizagem de máquina podem ter dificuldade de lidar com grandes volumes de dados
- ▶ Assim, torna-se útil obter uma amostra **representativa** dos dados para treinar o modelo
 - ▶ Os dados da amostra devem seguir a mesma distribuição dos dados originais (**qual?**)
- ▶ Diferentes amostras podem gerar modelos diferentes (mais sobre isso na aula sobre avaliação de modelos)

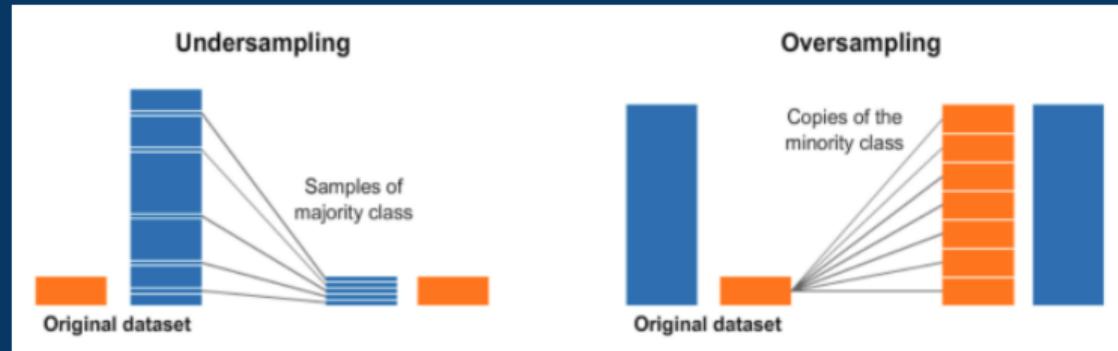


Balanceamento de dados

- ▶ Em certas aplicações (como na medicina), é comum que uma classe seja muito mais frequente do que outra
- ▶ Nesses casos, o modelo de AM pode aprender a “chutar” sempre a classe mais frequente
- ▶ Soluções:
 - ▶ Equalizar os tamanhos das classes
 - ▶ Subamostragem (*undersampling*)
 - ▶ Sobreamostragem (*oversampling*)
 - ▶ Classificação baseada em custos (mais sobre isso nos tópicos adicionais no final do curso)
 - ▶ Ajustar um modelo por classe



Balanceamento de dados



	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
4	4340	Claudia	21	F	52	Uniformes	37.6	1	PE	Doente
5	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
0	4201	Joao	28	M	79	Concentradas	38.0	2	SP	Doente

https://imbalanced-learn.readthedocs.io/en/stable/over_sampling.html#smote-adasyn



Limpeza dos dados

- ▶ Remove problemas relacionados à qualidade dos dados
- ▶ Dados ruidosos: erros de registro, variações de qualidade de sinal
 - ▶ Diferente de outliers
- ▶ Inconsistentes: contradizem valores de outros atributos do mesmo objeto
- ▶ Redundantes: dois ou mais objetos/atributos com os mesmos valores
- ▶ Incompletos (com ausência de valores)

Limpeza dos dados

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	NaN	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67.0	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92.0	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43.0	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52.0	Uniformes	NaN	1	PE	Doente
5	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
6	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
7	2301	Ana	22	F	72.0	Inexistentes	58.0	3	RJ	Doente
8	1322	Marta	19	F	87.0	Espalhadas	39.0	6	AM	Saudavel
9	3027	Paulo	34	M	67.0	Uniformes	38.4	2	GO	Saudavel



Dados incompletos

- Possibilidades de correção:
 - ▶ Eliminar instâncias/colunas com valores ausentes
 - ▶ Usar média/moda/mediana dos valores conhecidos
 - ▶ Criar um novo valor que indique o atributo tem valor faltante
 - ▶ Estimar a distribuição conjunta dos atributos para depois preencher os faltantes com os valores mais prováveis
 - ▶ Usar algoritmos capazes de lidar com dados ausentes

Dados inconsistentes

- ▶ Problemas na anotação dos dados podem resultar em atributos de entrada que não explicam o atributo alvo/classe

	identificador	nome	idade	sexo	peso	manchas	temperatura	internacoes	estado	diagnostico
0	4201	Joao	28	M	67	Concentradas	38.0	2	SP	Doente
1	3217	Maria	18	F	67	Inexistentes	39.5	4	MG	Saudavel
2	4039	Luiz	49	M	92	Espalhadas	38.0	2	RS	Doente
3	1920	Jose	18	M	43	Inexistentes	38.5	8	MG	Saudavel
4	4340	Claudia	21	F	52	Uniformes	38.5	1	PE	Doente
5	4340	Claudia	21	F	52	Uniformes	38.5	1	PE	Saudavel
6	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
7	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
8	2301	Ana	22	F	72	Inexistentes	58.0	3	RJ	Doente
9	1322	Marta	19	F	87	Espalhadas	39.0	6	AM	Saudavel
10	3027	Paulo	34	M	67	Uniformes	38.4	2	GO	Saudavel



Dados redundantes

- ▶ O mesmo atributo pode aparecer em dois formatos diferentes: idade X data de nascimento (string ou colunas numéricas)
- ▶ Atributos podem ser altamente correlacionados
 - ▶ Não há acréscimo de informação ao manter os dois
 - ▶ Mantém-se apenas um
 - ▶ Boa parte dos algoritmos de AM assume que não há correlação entre atributos

Outliers

- ▶ Dados que diferem bastante dos outros elementos do conjunto de dados ou de sua classe
- ▶ Podem ser retirados ou mantidos, caso deseje-se gerar modelos que modelam a sua existência
- ▶ Existem técnicas cujo objetivo é detectar outliers.

Transformação de dados

- ▶ Frequentemente é necessário transformar os tipos ou valores dos atributos para obter um melhor ajuste dos modelos
- ▶ Pode-se discretizar valores numéricos ou transformá-los em intervalos
- ▶ Pode-se transformar atributos categóricos com p categorias em p atributos binários
 - ▶ One-hot encoding, variáveis dummy
- ▶ E fazemos também a conhecida normalização, quando os atributos têm escalas muito diferentes

$$X_{novo} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Novos valores entre $[0, 1]$

$$Z = \frac{X - \mu}{\sigma}$$

Lida melhor com outliers



Aprendizagem de Máquina

Conceitos Fundamentais

Juliana Freitas Pires

juliana.freitas@academico.ufpb.br

www.de.ufpb.br

UFPB

 Departamento de
ESTATÍSTICA