Machine learning em saúde

Prof. Dr. Alexandre Chiavegatto Filho

Python para a análise de dados

KDNUGGETS

ANNUAL POLL

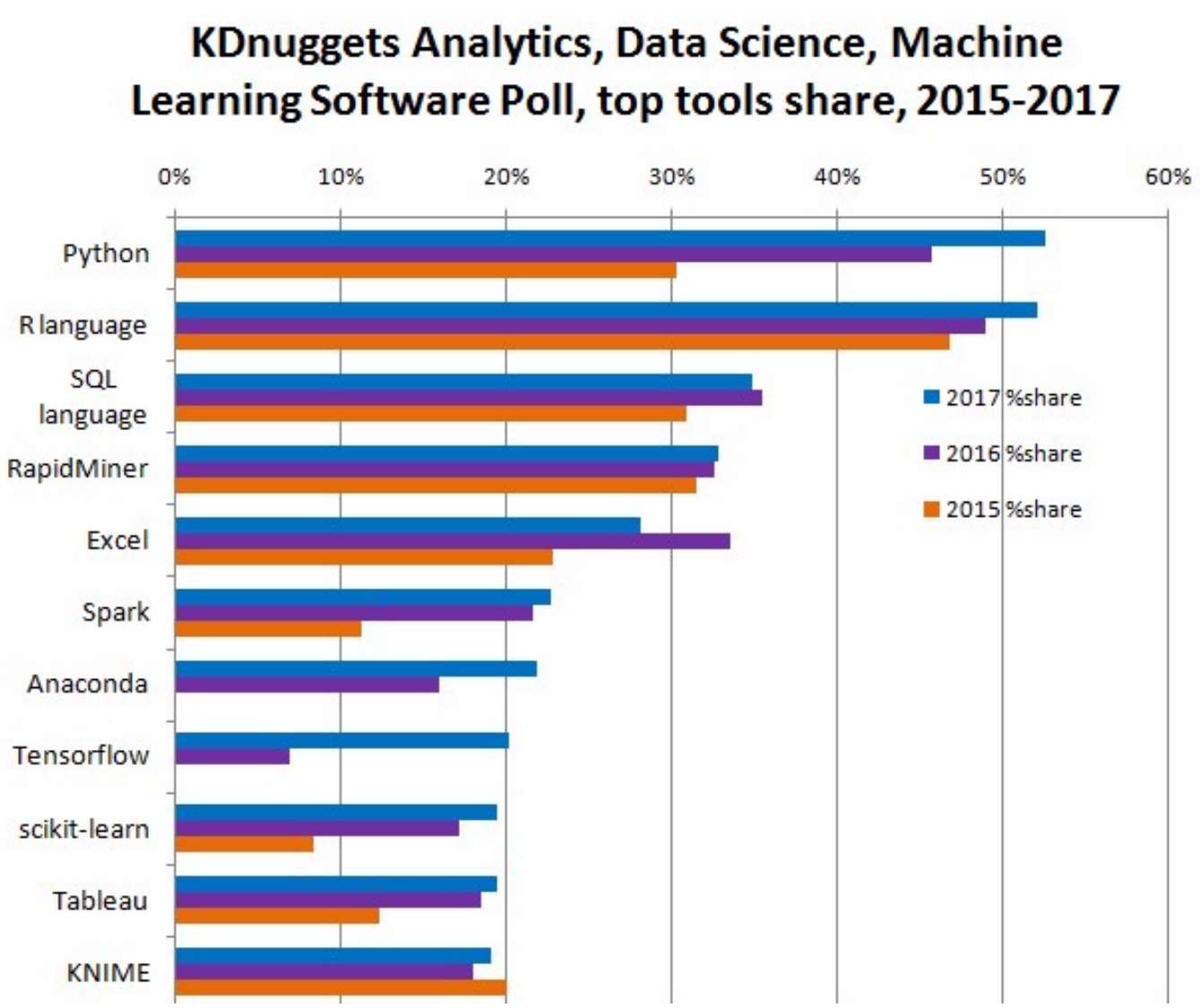
2017

2.900 cientistas de dados

"Quais softwares você usou para analisar dados nos últimos 12 meses?"

52,1%

python™ 52,6%.



Linguagem de programação

Muito fácil fazer alterações nas análises
(1 linha de código)
Facilita análises colaborativas
Garante reprodutibilidade dos resultados em novas amostras

Gratuito

Comunidade ativa de programadores

Foco na análise de dados

Linguagem de programação

Muito fácil fazer alterações nas análises
(1 linha de código)
Facilita análises colaborativas.
Garante reprodutibilidade dos resultados em novas amostras.

Gratuito

Comunidade ativa de programadores

Linguagem de programação geral

Em machine learning, bastante utilizado para deep learning.

Maioria dos pesquisadores da área usam o Python – maior disponibilidade de códigos e tutoriais.

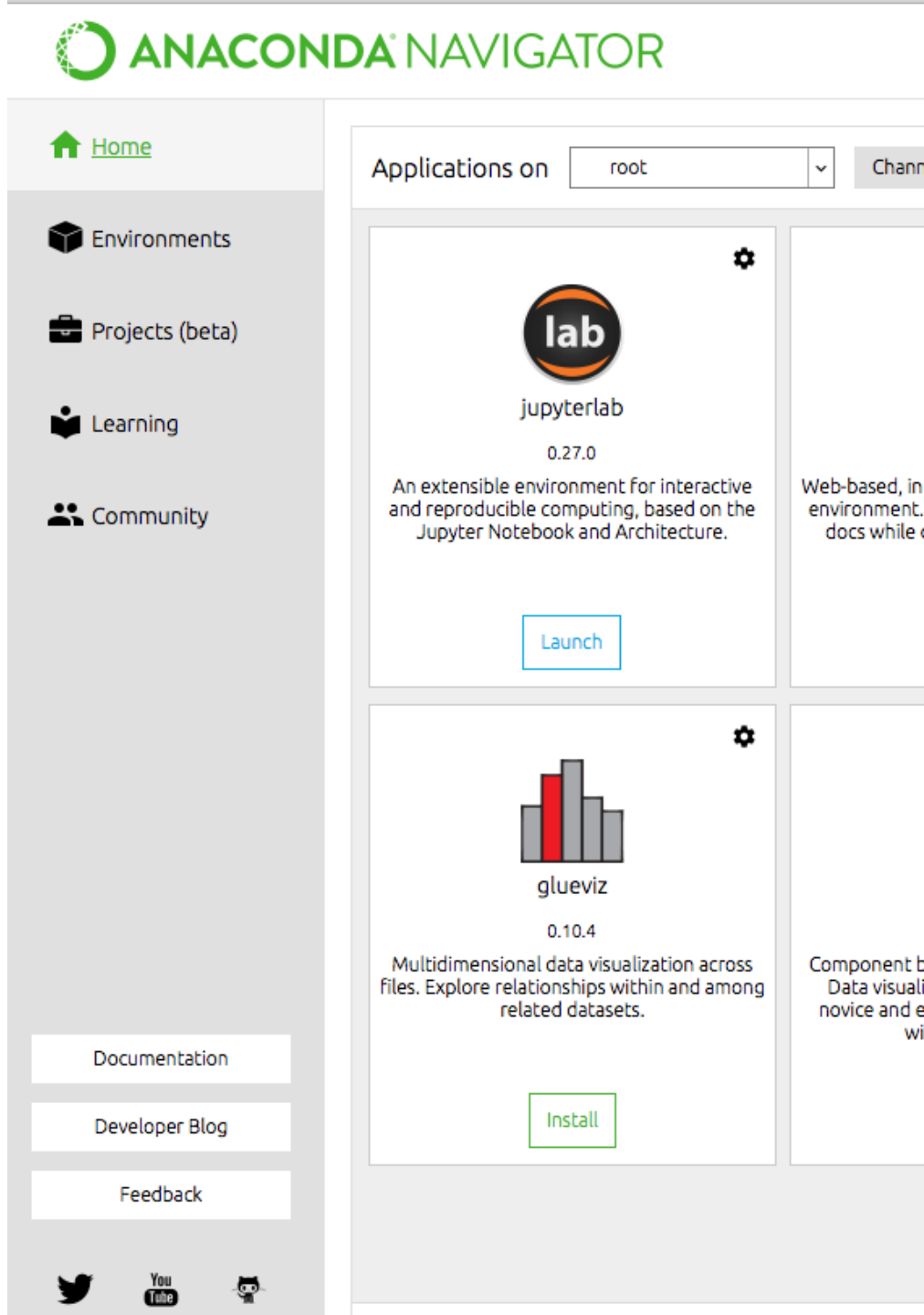
Porém deep learning (keras) também roda no R.

Praticamente empatado com o R no uso entre cientistas de dados.

ANACONDA.

- Inclui as principais bibliotecas para analisar dados no Python.
  - -2015: 125 bibliotecas.
  - -2016: 300 bibliotecas.
  - -2017: 720 bibliotecas.
  - -2018: 1000 bibliotecas.
  - -2019: 1400 bibliotecas.

- Inclui Jupyter Notebook: interface do Python para análise de dados.



QUATRO CATEGORIAS DE MACHINE LEARNING

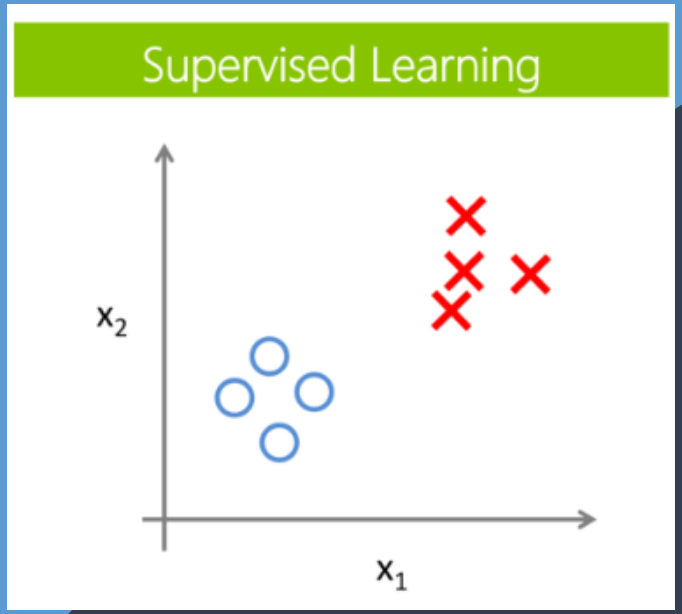
Objetivo: acertar um resultado que existe.

Treinar um modelo para obter a melhor performance preditiva possível para um *problema*

APRENDIZADO SUPERVISIONADO

Quando os dados incluídos para treinar o algoritmo incluem a solução desejada, ou rótulo (“label”).

RESPOSTA CERTA



Aprendizado supervisionado

Divididos em dois grandes grupos

CLASSIFICAÇÃO

Quando a variável a ser predita é categórica

Exemplo

Óbito em 5 anos, incidência de doença em 10 anos, etc.

REGRESSÃO

Quando a variável a ser predita é quantitativa

Exemplo

Quantos meses de vida a pessoa tem pela frente, qual será o seu IMC no próximo ano, etc.

Aprendizado não-supervisionado

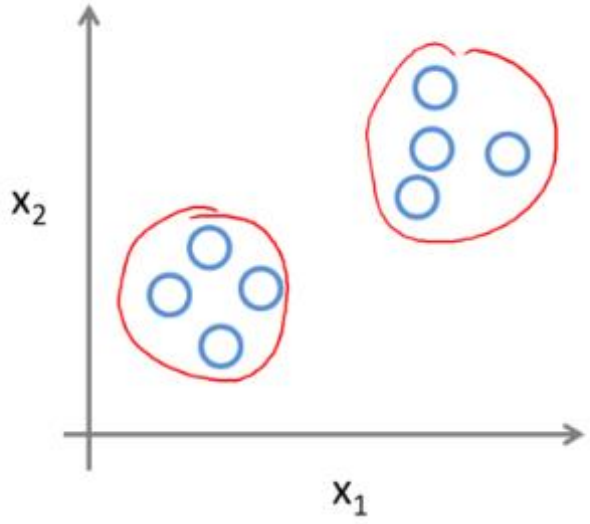
Não existe rótulo (“label”).

Algoritmo aprende sem uma resposta certa.

O objetivo é encontrar padrões nos dados.

Mais comuns: clustering (agrupamentos) e redução de dimensão (ACP).

APRENDIZADO NÃO-SUPERVISIONADO



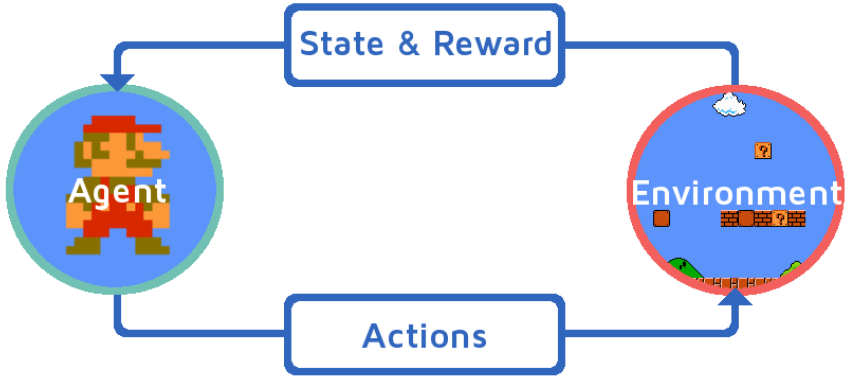
Aprendizado semi-supervisionado

Presença de alguns dados com rótulo e outros sem.

Identificação de fotos do Facebook: algoritmo identifica que a mesma pessoa está em várias fotos e só precisa de um rótulo.

Aprendizado por reforço

- Agente interage com um ambiente dinâmico.
- Feedbacks constantes em termos de premiações e punições.
- Jogos.



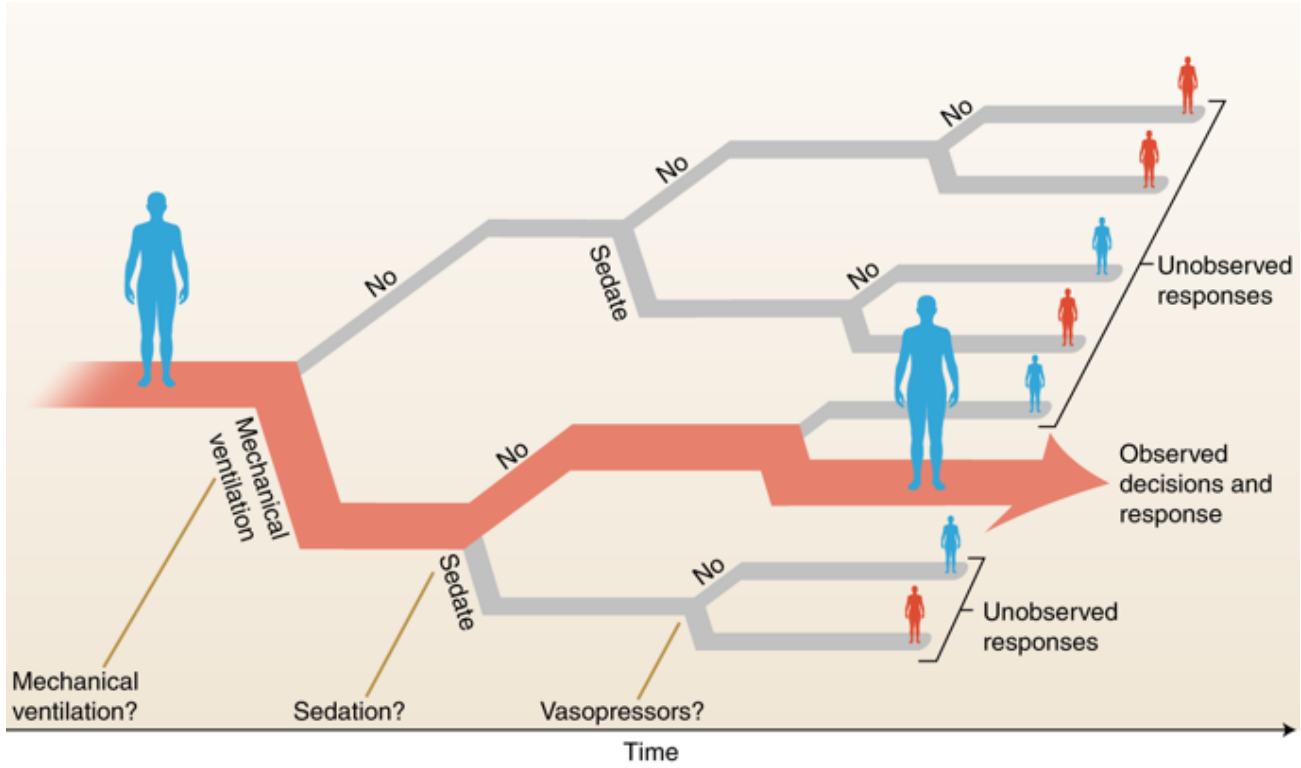
QUATRO CATEGORIAS DE MACHINE LEARNING

Aprendizado por reforço

Área promissora:

- Diferentes etapas do tratamento médico para identificar sequência ótima (ex: sepse).

Gottesman O, Johansson F, Komorowski M, Faisal A, Sontag D, Doshi-Velez F, Celi LA. Guidelines for reinforcement learning in healthcare. Nat Med 2019. 25(1)16-18.



PREDIÇÃO COM MACHINE LEARNING

- Trade-off entre predição e interpretação.

- Em predição estamos interessados em performance preditiva.

- Em inferência o interesse é entender a relação entre variáveis, normalmente como Y muda com uma alterações entre seus determinantes.

- Problema: modelos facilmente interpretáveis (regressões linear e logística, árvores de decisão) normalmente têm pior performance preditiva.

- Pensar bem sobre o objetivo da análise: é inferência ou predição?

PREDIÇÃO COM MACHINE LEARNING

Dados

Preferencialmente muitos e com boa qualidade (preenchidos corretamente e preditores fortes). Realizar o pré-processamento das variáveis.

Algoritmos

Inserir os dados no algoritmo de machine learning para aprender os parâmetros (regressão logística/linear) ou estruturas (árvores) que mapeiam os preditores aos resultados.

Testar no futuro

Inserir no algoritmo novos dados para testar a qualidade desse algoritmo para predizer dados futuros.

PREDIÇÃO COM MACHINE LEARNING

- Desenvolver algoritmos que façam boas predições em saúde.
- Principais razões pelas quais algoritmos às vezes não apresentam boa performance preditiva:
  - Extrapolação inadequada dos resultados.
  - Pré-processamento inadequado dos dados.
  - Sobreajuste (mais importante).
  - Validação inadequada da qualidade dos algoritmos.

Extrapolação inadequada

- Desenvolver os algoritmos para uma população e esperar que funcionam corretamente para outra diferente.
  - Importar algoritmos dos EUA/Europa: nossas características genéticas e socioeconômicas são muito diferentes.
  - Extrapolação para períodos diferentes (cuidado com doenças sazonais).

PRÉ-PROCESSAMENTO DOS DADOS

- Técnicas de pré-processamento de dados
  - Seleção das variáveis.
  - Vazamento de dados.
  - Padronização.
  - Redução de dimensão.
  - Colinearidade.
  - Valores missing.
  - One-hot encoding.

PRÉ-PROCESSAMENTO DOS DADOS

- Pré-selecionar variáveis que sejam preditoras plausíveis (bom senso do pesquisador).

- Coincidências acontecem em análises de big data e pode ser que o algoritmo dê muita importância para associações espúrias.

PRÉ-PROCESSAMENTO DOS DADOS

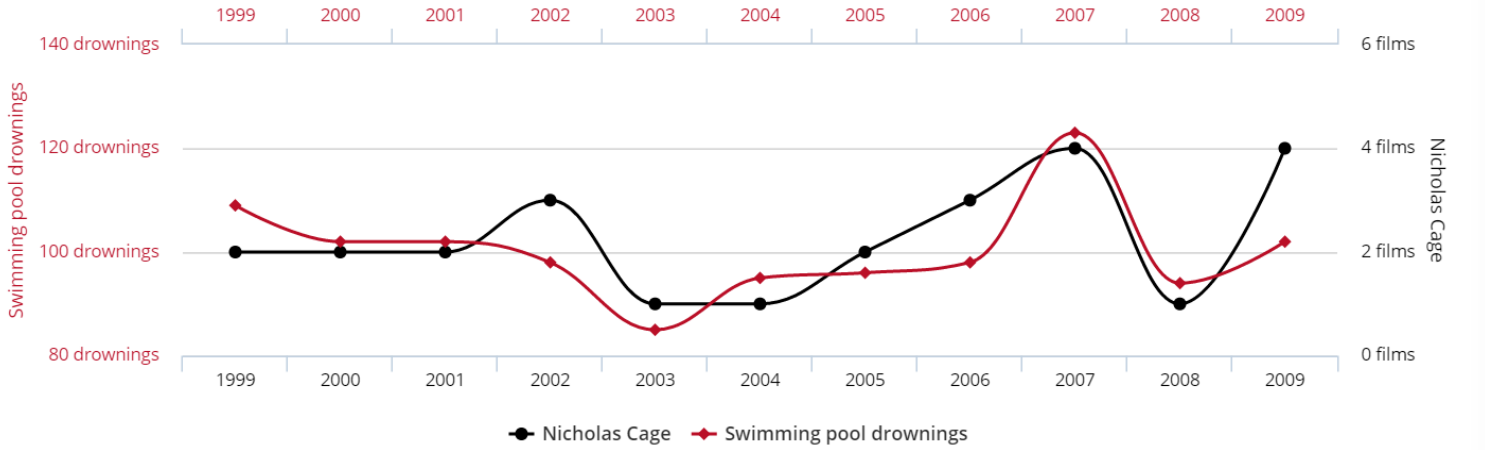
tylervigen.com

Number of people who drowned by falling into a pool

correlates with

Filmas Nicolas Cage appeared in

Correlation: 66,6% (r=0,666004)



PRÉ-PROCESSAMENTO DOS DADOS

Cuidado com vazamento de informação (“data leakage”).

- Acontece quando os dados de treino apresentam informação escondida que faz com que o modelo aprenda padrões que não são do seu interesse.

- Uma variável preditora tem escondida o resultado certo:
    - Não é a variável que está predizendo o desfecho, mas o desfecho que está predizendo ela.

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo

Incluir o número identificador do paciente como variável preditora

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo	Problema
Incluir o número identificador do paciente como variável preditora	Se pacientes de hospital especializado em câncer tiverem números semelhantes. Se o objetivo for predizer câncer, algoritmo irá dar maior probabilidade a esses pacientes. Esse algoritmo aprendeu algo interessante para o sistema de saúde?

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo	Problema	Motivo
Incluir o número identificador do paciente como variável preditora	Se pacientes de hospital especializado em câncer tiverem números semelhantes. Se o objetivo for predizer câncer, algoritmo irá dar maior probabilidade a esses pacientes. Esse algoritmo aprendeu algo interessante para o sistema de saúde?	Motivo pelo qual os dados e os algoritmos de machine learning precisam ser abertos. Watson prediz bem: mas é informação útil ou vazamento?

PADRONIZAÇÃO

- A escala das variáveis pode afetar muito a qualidade das predições.

- Alguns algoritmos darão preferência para utilizar variáveis com valores muito alto.

PADRONIZAÇÃO

- Padronizar as variáveis contínuas para todas terem média de 0 e desvio-padrão de 1.

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Ou seja, é feita a subtração da média e a divisão pelo desvio padrão dos valores da variável.

REDUÇÃO DE DIMENSÃO

- Quanto maior a dimensão dos dados (número de variáveis) maior o risco de o algoritmo encontrar e utilizar associações espúrias.

REDUÇÃO DE DIMENSÃO

- Análise de Componentes Principais

Técnica de aprendizado não supervisionado.

O objetivo é encontrar combinações lineares das variáveis preditoras que incluam a maior quantidade possível da variância original.

O primeiro componente principal irá preservar a maior combinação linear possível dos dados, o segundo a maior combinação linear possível não correlacionada com o primeiro componente, etc.

VARIÁVEIS COLINEARES

Uma das razões pela qual a ACP é tão utilizada, é o fato de que cria componentes principais não correlacionados.

- Na prática, alguns algoritmos conseguem melhor performance preditiva com variáveis com baixa correlação.

VARIÁVEIS COLINEARES

Uma outra forma de diminuir a presença de variáveis com alta correlação é excluí-las.

- Variáveis colineares trazem informação redundante (tempo perdido).
- Além disso, aumentam a instabilidade dos modelos.
- Estabelecer um limite de correlação com alguma outra variável (0,75 a 0,90).

VARIÁVEIS MISSING

É importante entender por que valores de uma variável estão faltantes.

Motivo sistemático → INFORMAÇÃO PREDITIVA.

Grande diferença em relação a estudos de inferência, em que valores missing devem ser evitados.

Não conseguiu responder a uma pergunta sobre o seu passado → INFORMAÇÃO PREDITIVA.

Pode ajudar na predição de problemas cognitivos graves no futuro

Em variáveis categóricas adicionar uma categoria para missing.

Imputação com machine learning.

ONE-HOT ENCODING

Alguns algoritmos têm dificuldade em entender variáveis que têm mais do que uma categoria.

Acham que é uma variável contínua (0, 1, 2, 3...) → porém não têm significado contínuo.

A solução é transformar todas as categorias em uma variável diferente de valores 0 e 1 (one-hot encoding).

Variável com n categorias → criadas n variáveis.

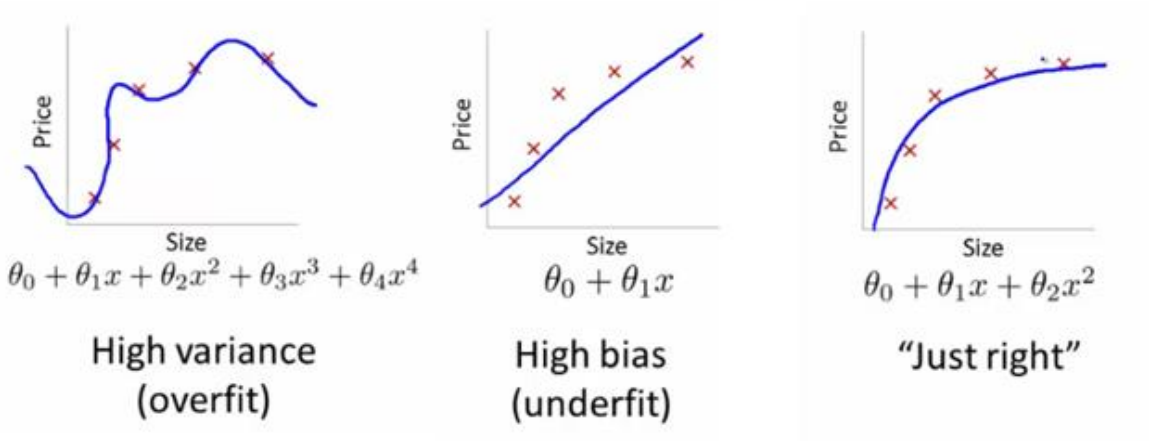
ONE-HOT ENCODING

Pode trazer problemas em alguns modelos, como na regressão linear.

Solução: criar dummies.

n-1 variáveis (deixar a mais frequente como categoria de referência).

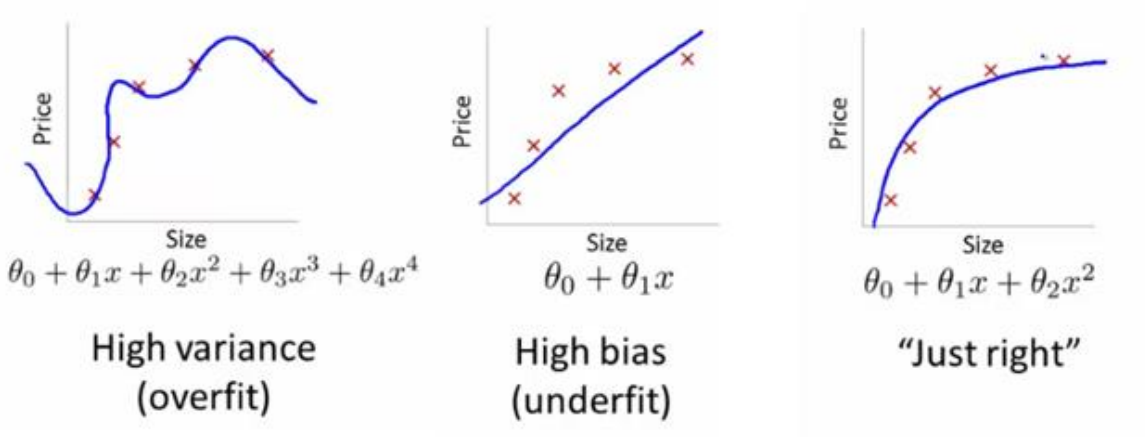
SOBREAJUSTE



Principal problema de machine learning

Modelos muito complexos:

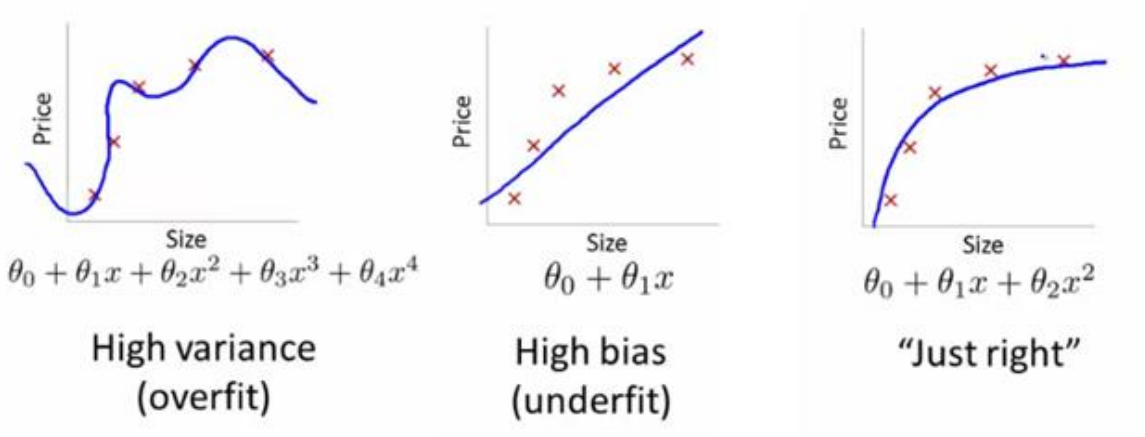
- Funcionam perfeitamente para a amostra em questão, mas não muito bem para amostras futuras.
  - Dados influenciados por fatores aleatórios e erros de medida.



Principal problema de machine learning.

Tradeoff entre viés e variância:

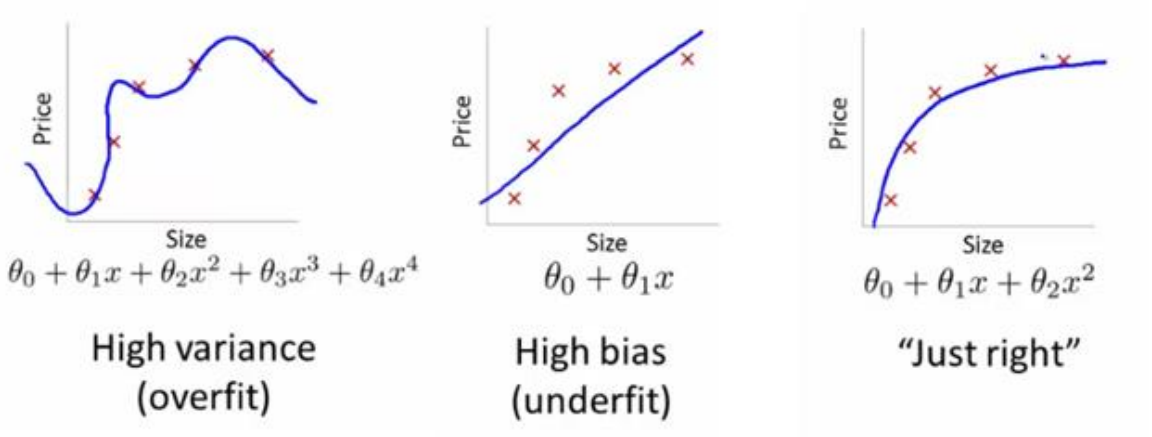
- Viés: erro gerado pelo uso de modelos para dados reais.
- Variância: quando pequenas mudanças nos dados levam a uma mudança muito grande nos parâmetros



Tradeoff entre viés e variância:

Modelo com alta variância e pouco viés

- 2 variáveis: linha que passa exatamente por todos os pontos.
- Se ajusta perfeitamente aos dados atuais, mas não aos futuros.



Tradeoff entre viés e variância:

Modelo com baixa variância e alto viés

- 2 variáveis: linha reta para associação não-linear.
- Modelo simples, com baixo poder preditivo.

Como avaliar se o seu modelo está com sobreajuste?

- Avaliar a performance preditiva do modelo em dados que não foram utilizados para definir o modelo.
    - Se a performance preditiva cair muito com os novos dados: o modelo tem sobreajuste.
    - É muito fácil ter boa predição nos dados que foram utilizados para definir o modelo: é só tornar o modelo muito complexo.

Soluções

Utilizar dados do período seguinte.

- Exemplo: treinar o modelo em dados de 2016 e avaliar sua performance em dados de 2017.
- Problema: na maioria das vezes, os dados são coletados num mesmo período.

Separar os dados aleatoriamente em treino e teste.

Dados de “treino” (70-80%) são usados para definir o modelo e dados de “teste” (20-30%) são usados **uma única vez** para analisar a performance preditiva final dos modelos.

Para análises em que o desfecho a ser predito é uma categoria:

- Amostragem estratificada entre treino e teste.
- Manter mesma proporção do desfecho de interesse nos dois grupos.

Alguns algoritmos na prática conseguem melhor performance com distribuição igual entre as categorias: desfecho binário com 50% cada.

Soluções:

- Down-sampling: selecionar amostra da classe mais frequente até se igualar à menos frequente.
- Up-sampling: amostragem com reposição da classe menos frequente até se igualar à mais frequente.
- SMOTE: combinação de down e up-sampling.

O QUE SIGNIFICA “DEFINIR” O MODELO?

- Estabelecer os parâmetros (definidos automaticamente) e os hiperparâmetros (definidos pelo pesquisador).

- Hiperparâmetros são em geral regularizadores: ou seja, tentam controlar a complexidade dos modelos (para evitar o sobreajuste).

COMO SELECIONAR OS VALORES DOS HIPERPARÂMETROS?

Pela análise da melhora da performance preditiva.

Problema: dados de teste só podem ser usados **uma única vez**, para a seleção do melhor algoritmo.

SOLUÇÕES

1

Se tiver muitas observações: separar um terceiro grupo (validação) para ajuste dos hiperparâmetros.

COMO SELECIONAR OS VALORES DOS HIPERPARÂMETROS?

Pela análise da melhora da performance preditiva.

Problema: dados de teste só podem ser usados **uma única vez**, para a seleção do melhor algoritmo.

SOLUÇÕES

2

Se tiver poucas observações: selecionar os hiperparâmetros nos dados de treino.

- Problema: performance dos modelos deve sempre ser testada em dados que o algoritmo nunca viu (evitar sobreajuste).
- Solução: validação cruzada.

VALIDAÇÃO CRUZADA DE 10 PARTES (10-FOLD)

- Dividir os dados em dez partes iguais e utilizar nove delas para treinar o algoritmo com um hiperparâmetro e a outra parte para testar a sua predição.