Introdução a Machine Learning (aprendizado de máquina) 1

Semana 1

O que é machine learning (aprendizado de máquina) ?

Não há um consenso sobre a definição. Aqui vão duas tentativas:

- “O campo de estudo que permite que computadores aprendam sem ser explicitamente programados”.
- “Um programa de computador aprende uma tarefa T com a experiência E , medido pela performance P se a performance em T, melhora com respeito a P com o aumento de experiência E.”

Como explicar a diferença entre cães e gatos? `

Duas abordagens:

1. Programação
2. Machine Learning

Outros Exemplos:

1. Filtros de spam nos emails
2. Reconhecimento de dígitos
3. Tradução automática
4. Segmentação dos clientes
5. Carro que dirige sózinho
6. Reconhecimento de voz
7. Escolher ações no mercado financeiro
8. Estimar o tempo de chegada do seu Uber

No primeiro exemplo, a tarefa T é classificar os emails como spam, a experiência E é o número de emails que o computador observa, e a performance P poderia ser a proporção de emails que o computador classifica corretamente.

Sobre esse curso

Semana 1

Métodos diferentes

3 métodos de ensino:

a. Cima para baixo: Um pouco de matemática e intuição sobre os algoritmos
b. Baixo para cima: Exemplos de códigos
c. Conselhos práticos sobre como criar algoritmos de ML

O que vamos aprender nesse curso

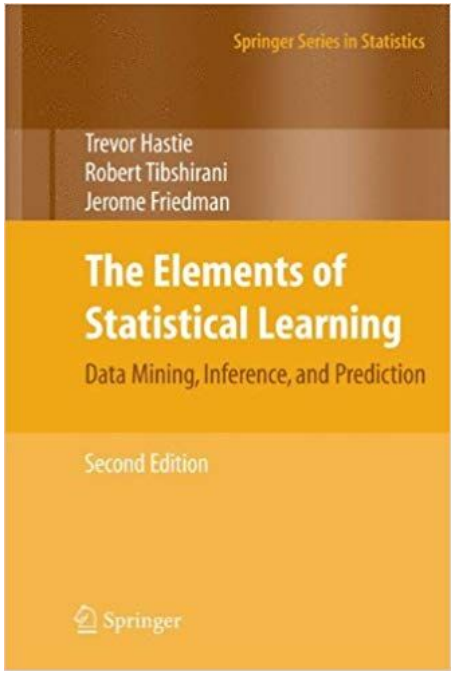
1. Introdução a machine learning
2. Modelos, função objetivo, método do gradiente
3. Regressão Linear
4. Regressão Logística
5. Redes Neurais
6. Introdução a deep learning
7. Árvores de decisão e florestas aleatórias
8. Aprendizado não supervisionado e sistemas de recomendação

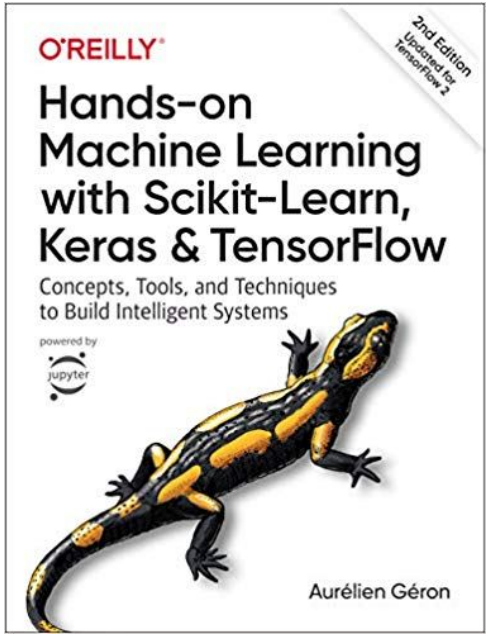
Quem é o público alvo

Pessoas podem se beneficiar de formas diferentes do curso

1. Estudantes verão um pouco da matemática e um pouco da probabilidade por trás dos algoritmos.
2. Programadores vão ver quão fácil é escrever código usando as ferramentas de aprendizado de máquina.
3. Profissionais podem receber conselhos práticos sobre como construir algoritmos.

Outras fontes





Um pouco sobre mim - Tiago

PUC RIO

UCLA BRUINS

airbnb

UBER

GOODWATER CAPITAL

3 formas de aprendizado de máquina

Aula 1

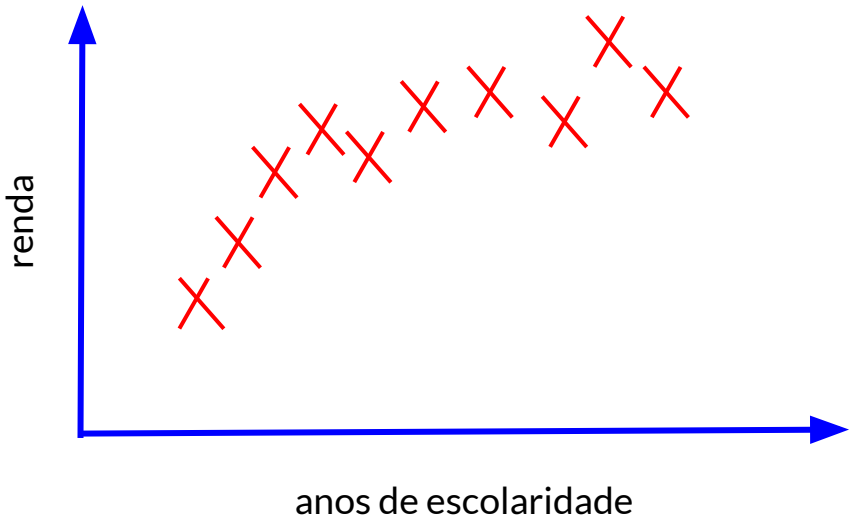
3 formas de machine learning

1. Aprendizado supervisionado
2. Aprendizado não supervisionado
3. Aprendizado por reforço (Reinforcement Learning)

Exemplo de aprendizado supervisionado: regressão

Exemplo: previsão de renda.

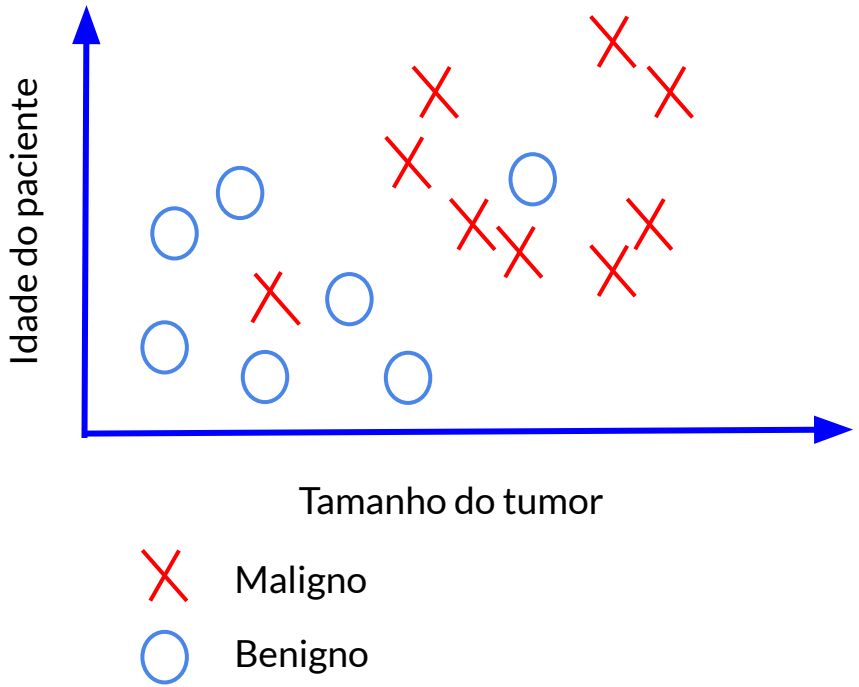
- Nós sabemos a verdadeira renda de cada indivíduo
- Para cada indivíduo, temos as características X e os rótulos Y.
- Nesta caso Y é uma variável contínua, ou seja, Y pode assumir infinitos valores distintos.
- Esse é um exemplo de regressão.



Exemplo de aprendizado supervisionado: classificação

Exemplo: O tumor é maligno

- Novamente, nós sabemos o rótulo correto.
- Temos 2 características e 1 rótulo Y.
- Y é uma variável binária, só assume valores 1 (tumor maligno) ou 0 (benigno).
- Esse é um problema de classificação, o rótulo Y assume um número finito de valores.



Algoritmos de aprendizado supervisionado

Alguns dos algoritmos mais importantes de aprendizado supervisionado:

1. Regressão linear
2. Regressão logística
3. Multinomial Logit
4. Regressão Lasso e Ridge
5. K-vizinhos (K-nearest neighbors)
6. Árvores de decisão
7. Florestas Aleatórias
8. Redes Neurais

Aprendizado não supervisionado

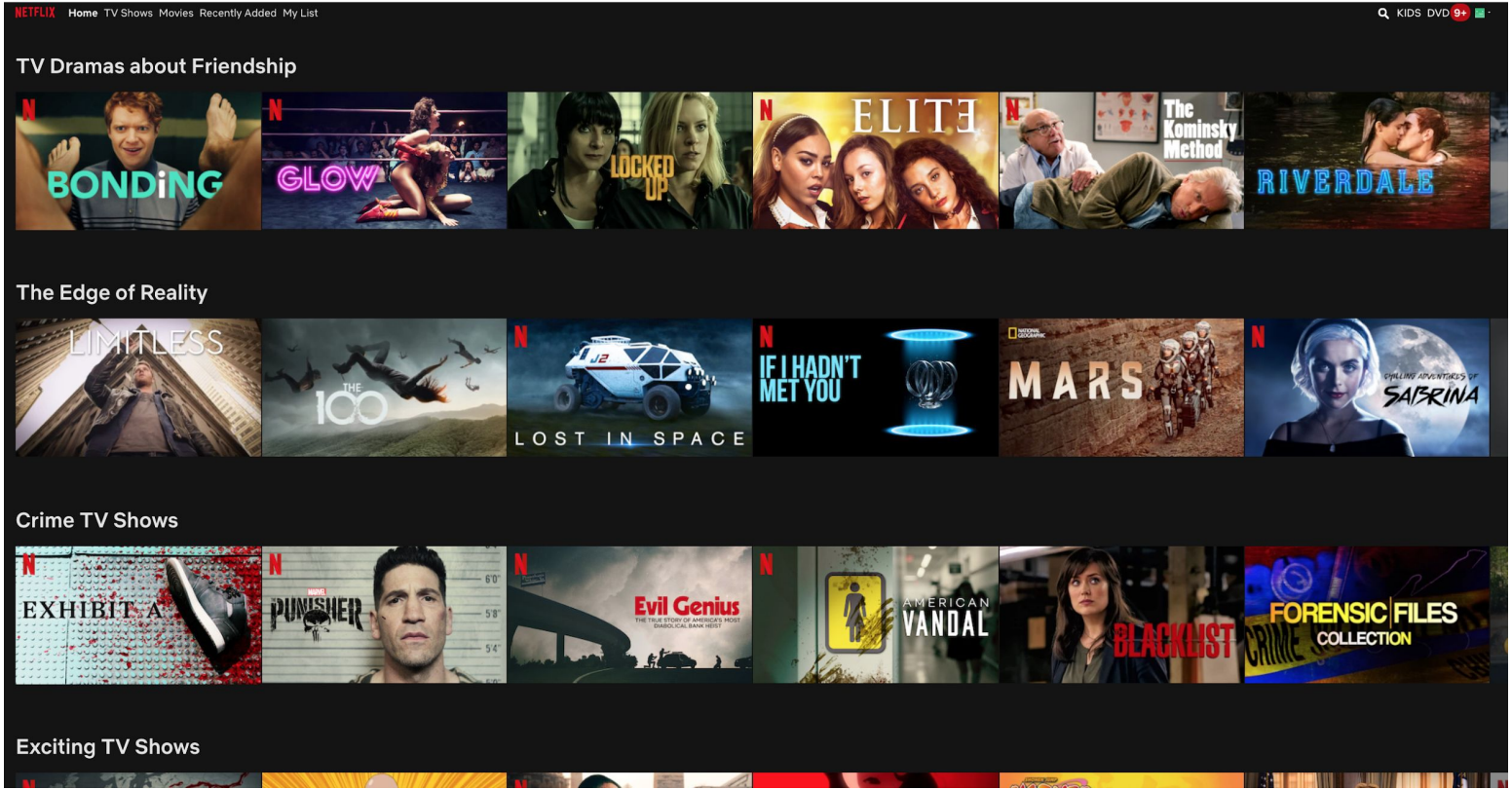
Aula 1

Aprendizado não supervisionado

Exemplo: Segmentação dos clientes

- Não há rótulos (Y), apenas características (X).
- Muitas vezes você tem que decidir qual é a pergunta relevante.
- Como dividir os clientes em segmentos diferentes?

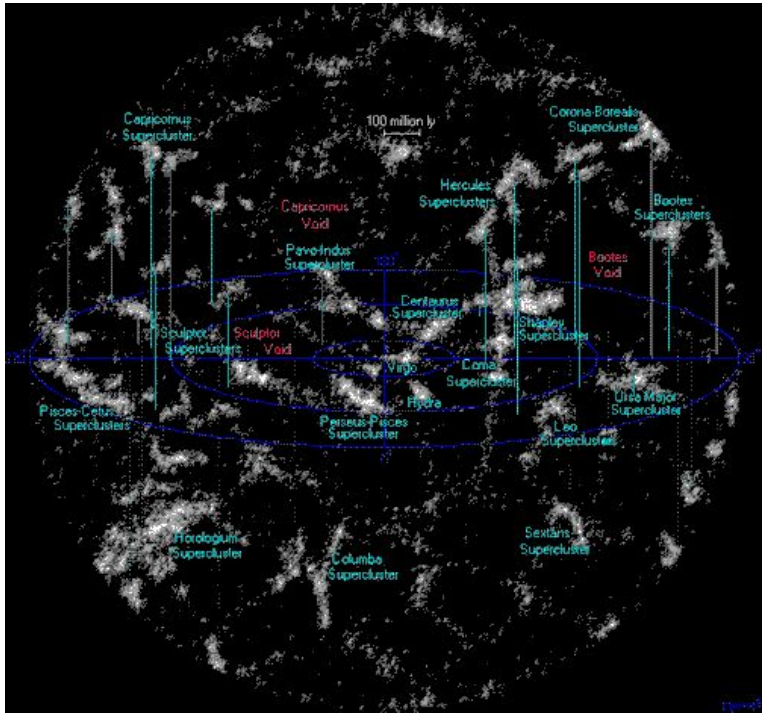
Categorias de shows no Netflix



Aplicações de aprendizado não supervisionado

Algumas aplicações:

1. Segmentação de clientes
2. Criar categorias de filmes
3. Detectar fraudes e anomalias
4. Organizar grupos de servidores
5. Agrupar notícas na internet
6. Análise do gráfico de redes sociais
7. Encontrar parentes usando genoma
8. Análise de dados astronômicos



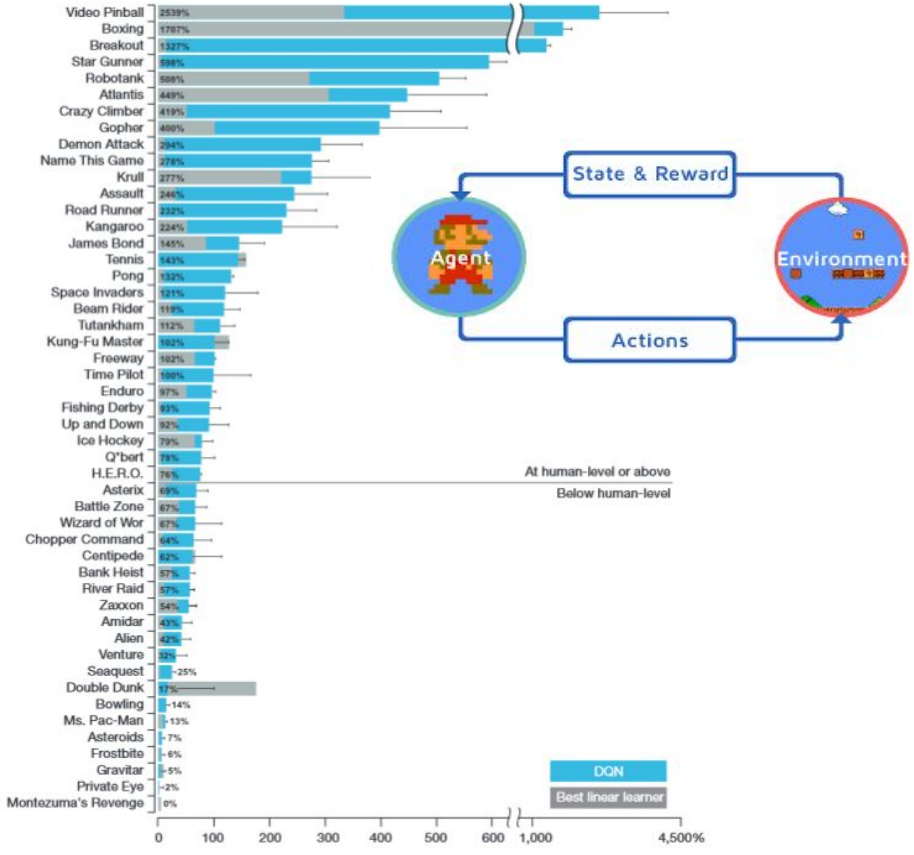
Algoritmos de aprendizado não supervisionado

Alguns dos algoritmos mais importantes de aprendizado não supervisionado:

1. K-médias (k-means)
2. Misturas Gaussianas
3. Autocodicadores com redes neurais (auto-encoders)

Aprendizado por reforço (Reinforcement Learning)

- Vários avanços em inteligência artificial vem de aprendizado por reforço.
- Alpha Go, records em video games e mercado financeiro são algumas aplicações
- Útil quando há um ambiente onde podemos simular o resultado muitas vezes e há um objetivo bem definido.
- Não vamos cobrir nesse curso.
- Aulas do David Silver no Youtube.



Desafios comuns no aprendizado de máquina

Semana 1

Desafios do aprendizado de máquinas

1. Pouca quantidade de dados para treinar o modelo
2. Dados não representativos e viés de seleção
3. Qualidade ruim dos dados
4. Características irrelevantes
5. Super-adequação (overfitting) dos dados
6. Sub-adequação (underfitting) dos dados

Pouca quantidade de dados

- A eficácia irracionais dos dados
- Precisamos pensar com cuidados sobre a dicotomia entre algorítimos e dados

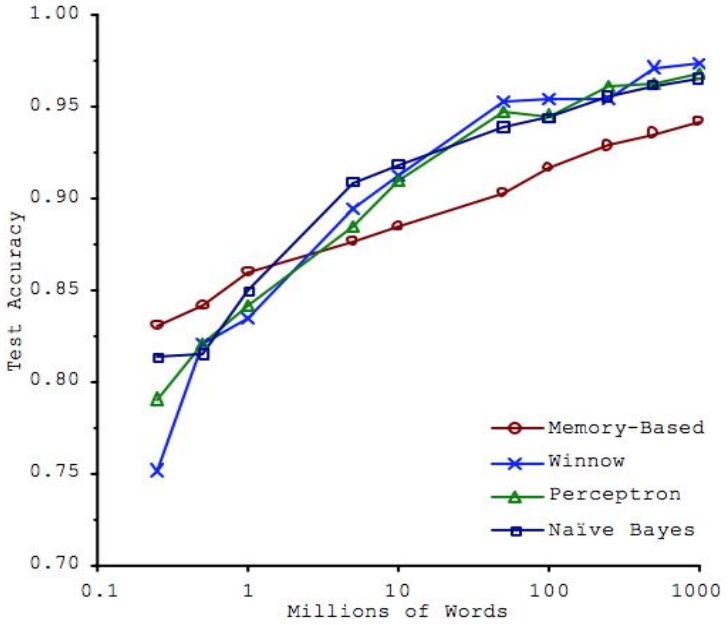
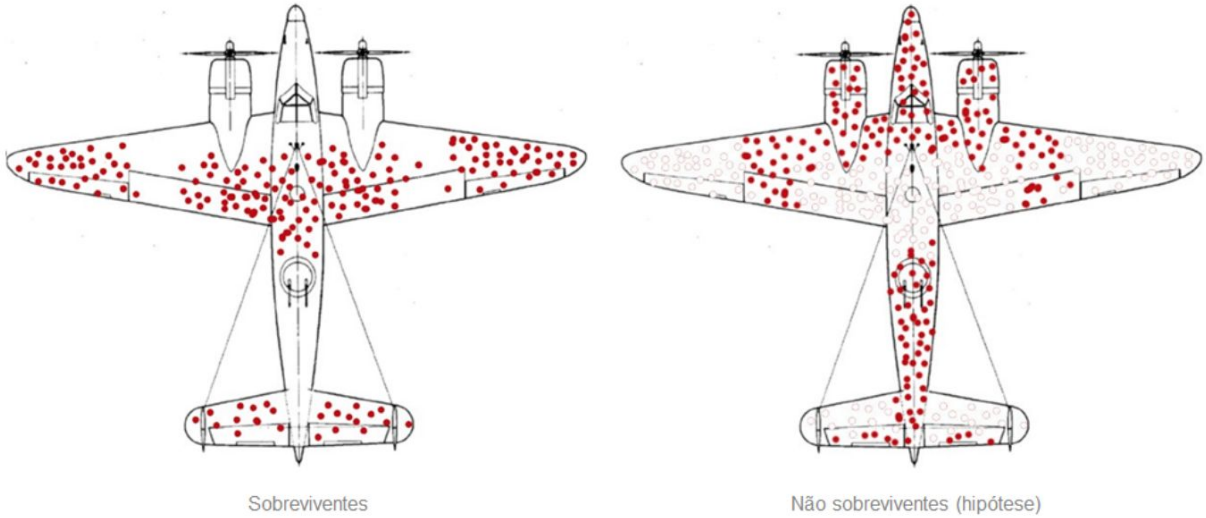


Figure 1. Learning Curves for Confusion Set Disambiguation

Base de dados não representativa

- Seu modelo não vai extrapolar bem se se a base de dados que você usou para treinar o algoritmo for diferente da base de dados na qual você quer testar o modelo

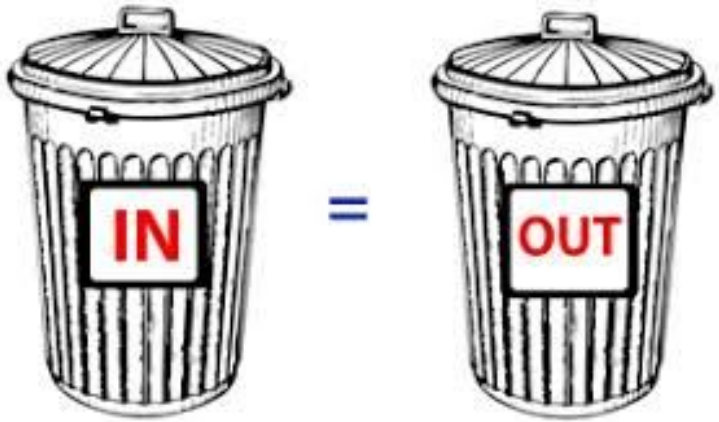
Viés de seleção



Pesquisas previram que Landon ganharia a eleição de 1936 nos EUA com 57% dos votos. FDR ganhou com 62%
NYT artigo mostrando como as pesquisas erraram na eleição do Trump

Qualidade ruim dos dados

- Lixo dentro, lixo fora (garbage in, garbage out)
- Como melhorar a qualidade dos dados?
  - Limpar a base de dados
  - Remover dados extremos (outliers)
  - Imputação de dados



Características Irrelevantes

- Características irrelevantes aumentam a proporção de lixo sobre a quantidade dados úteis
- Engenharia de características: permite criar as características mais informativas a partir da sua base de dados
- Extração de características: permite criar novas características através de combinações de características existentes. Exemplo é o algoritmo de análise de componentes principais (principal component analysis).

Sobre-adequação e sub-adequação

- Sobre-adequação (overfitting) ocorre quando o modelo é muito complexo, ele explica o conjunto de treinamento bem demais e não extrapola bem para outras bases de dados.
- Sub-adequação (underfitting) ocorre quando o modelo é muito simples e nem mesmo explica bem o conjunto de treinamento.

… to the code!