UEA
UNIVERSIDADE
DO ESTADO DO
AMAZONAS

Introdução à *Machine Learning*

Elloá B. Guedes
ebgcosta@uea.edu.br
www.elloaguedes.com

Encontro Regional de Pesquisa Operacional do Norte (ERPO 2018)

Apresentação

- **Elloá B. Guedes**

- Doutora em Ciência da Computação
- Atua na EST/UEA desde 2013
- Líder do Laboratório de Sistemas Inteligentes

- *Machine Learning*
- *Deep Learning*
- Entusiasta Python

Material do minicurso

https://github.com/elloa/erpo2018

Motivação


1
01
100
110
001
0110110010
01100110101 01
010101001 0111
001101010101101100
110
1101100

Motivação

- “**Dados** são o novo petróleo” (Humby,2006)

Motivação

- “**Dados** são o novo petróleo” (Humby,2006)

- É preciso encontrar, extrair, refinar, distribuir e monetizar
- Quantidades massivas de dados
- Fenômeno: Big Data

- Dados só possuem valor se puderem fornecer **insights**

Motivação

Machine Learning

"É o estudo sistemático de algoritmos e sistemas que melhoram o seu conhecimento ou performance com a experiência."

– Flach, P. Machine Learning, 2012. Cambridge University Press.

Machine Learning – Histórico

- Arthur Samuel, 1959, IBM
- Algoritmo jogador de Damas
- Primeira versão: equação de pontuação baseado em quantidade de peças e suas posições
- Segunda versão: melhorar os coeficientes da equação a partir de jogos
- Computador jogando contra si mesmo milhares de vezes
- Meados de 1970: performance comparável a de um amador

Machine Learning – Histórico

- Samuel: programa que melhorava a si mesmo a partir da experiência
- Nascimento do *Machine Learning*

- Não há fronteira clara entre IA e *Machine Learning*
- ML é uma forma de IA
- IA é mais abrangente
- ML: corpo de conhecimento, métodos e técnicas

Machine Learning – Ferramental

Machine Learning – Aplicações

- Detecção de fraude
- Recomendação de produtos
- Diagnóstico médico
- Análise de sentimentos
- Monitoramento de tempo real
- Milhares de outras!

- Descoberta de relações não-triviais
- Parábolas: fralda e cerveja

Classificando as flores Iris

- Edgard Anderson, Botânico
- 1935, Quebec, Canadá
- Estudo das flores Íris
- Sir Ronald Fisher, 1936
- Análise de discriminantes lineares

Classificando as flores Iris

- Catálogo de 150 flores Íris
- Largura e comprimento da pétala e da sépala
- Classificação correspondente
- Dataset no github do mini-curso

Mão na massa!

Machine Learning

Processo de Aprendizagem de Máquina

1. Coleta e preparação de dados
2. Seleção de características
3. (sem spoilers ainda!)

Machine Learning

- Desafio **cats vs dogs**
- Kaggle, 2007
- 25 mil imagens
- 12.5 mil imagens para avaliação

Machine Learning

- Abordagem algorítmica
  - Programar todos os detalhes de como diferenciar gatos e cachorros
  - Texturas, cores, formas geométricas
  - Quantidade inviável de regras a serem capturadas
  - E se você esqueceu os gatos Sphynx ou os cachorros Komondor?

Machine Learning

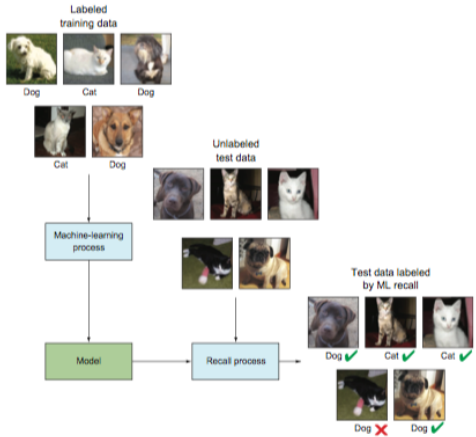
- Abordagem algorítmica
  - Programar todos os detalhes de como diferenciar gatos e cachorros
  - Texturas, cores, formas geométricas
  - Quantidade inviável de regras a serem capturadas
  - E se você esqueceu os gatos Sphynx ou os cachorros Komondor?

- Humanos cometem cerca de 7% de erros nesta tarefa

Machine Learning

- Abordagem *Machine Learning*
  - Análoga ao aprendizado de uma criança
  - Exemplos permitem o aprendizado de padrões
  - Capacidade de generalizar

Machine Learning



Machine Learning

- Abordagem *Machine Learning*
  - Análoga ao aprendizado de uma criança
  - Exemplos permitem o aprendizado de padrões
  - Capacidade de generalizar

- Modelos de *Machine Learning* para cats vs dogs
- 98.914% de acertos nos exemplos de avaliação

Machine Learning

- Precisamos organizar nossos dados

- Uma porção para fornecer experiência
- Outra porção para testar a capacidade de generalização do modelo

- Dados de treinamento e de testes

Machine Learning

- Vamos **particionar** os dados disponíveis

- 70% dos dados disponíveis para treinamento
- 30% dos dados para testes

- Randomizar ao particionar
- *Holdout cross-validation*

Machine Learning

Particionando os dados

```
from sklearn.model_selection import train_test_split
(...)
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.30)
```

Machine Learning

- Agora vamos fornecer os exemplos disponíveis ao modelo de aprendizado de máquina
- Aquisição de experiência!
- Captura de padrões nos dados

Machine Learning

Processo de Aprendizagem de Máquina

1. Coleta e preparação de dados
2. Seleção de características
3. **Escolha do algoritmo de aprendizado**

Escolha do Algoritmo

- **Há uma grande quantidade de algoritmos disponíveis**

Escolha do Algoritmo

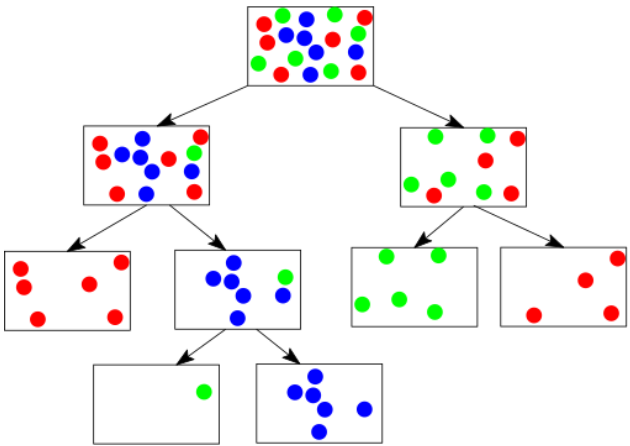
- **Há uma grande quantidade de algoritmos disponíveis**

- Características matemáticas
  - Regressão linear, regressão polinomial, etc.
- Inspirados no cérebro humano
  - Redes neurais artificiais
- Baseados em regras
  - Árvores de decisão
- Baseados na vizinhança
  - *k*-vizinhos mais próximos

Escolha do Algoritmo

- Vamos começar com **árvores de decisão**
- **Ideia geral**: Construir uma sequência de perguntas binárias a partir dos atributos preditores que permite descobrir a classe correspondente
- Menor número de perguntas que zera a incerteza sobre a classificação.

Escolha do Algoritmo



Escolha do Algoritmo

Criação do Modelo

```
from sklearn.tree import DecisionTreeClassifier
(...) arv = DecisionTreeClassifier()
```

Machine Learning

Processo de Aprendizagem de Máquina

1. Coleta e preparação de dados
2. Seleção de características
3. Escolha do algoritmo de aprendizado
4. Definição dos parâmetros
5. Treinamento
   - Fornecer os dados de treinamento para que o modelo adquira experiência

Treinamento do Modelo

Treinamento do Modelo

```
arv.fit(X_train,Y_train)
```

Machine Learning

Processo de Aprendizagem de Máquina

1. Coleta e preparação de dados
2. Seleção de características
3. Escolha do algoritmo de aprendizado
4. Definição dos parâmetros
5. Treinamento
6. Teste
   - Avaliação da performance
   - Coleta de métricas de desempenho
   - Comparação com outros modelos

Teste do Modelo

- **Vamos fazer previsões**

Teste do Modelo

- **Vamos fazer previsões**
- Importante: comparar com o gabarito

Teste do Modelo

- **Métricas de performance**
- Problema de classificação
- Acurácia
- Precisão
- Revocação
- F-Score

Teste do Modelo

Obtendo métricas de desempenho

```
from sklearn.metrics import accuracy_score
(...)
acc = accuracy_score(Y_predito,Y_test)
```

Comparando modelos

- Vamos testar outro algoritmo de Machine Learning
- *k*-vizinhos mais próximos

- Ideia geral: olhar para os *k* vizinhos mais próximos
- Valor previsto é a média dos valores dos vizinhos

- Vamos escolher $k = 5$ vizinhos
- Utilização de parâmetros no modelo

Comparando modelos

Algoritmo *k*-vizinhos mais próximos

```
from sklearn.neighbors import KNeighborsClassifier
(...)
# Inicializacao do modelo
kviz = KNeighborsClassifier(n_neighbors=3)
# Treinamento
kviz.fit(X_train,Y_train)
# Resultados para conjunto de testes
results = kviz.predict(X_test)
```

Comparando modelos

- Obtenha as métricas de desempenho para este modelo

Comparando modelos

- Obtenha as métricas de desempenho para este modelo
- Dentre os dois modelos considerados, qual o melhor para o problema em questão?

Comparando modelos

- Obtenha as métricas de desempenho para este modelo
- Dentre os dois modelos considerados, qual o melhor para o problema em questão?
- Precisamos ter cuidado para prevenir **overfitting**

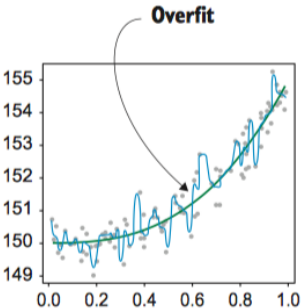
Overfitting

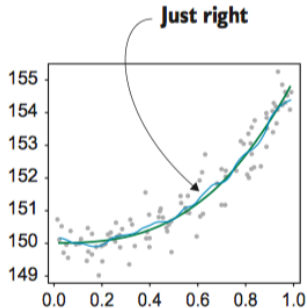
- ***Overfitting***: superajustamento do modelo aos dados de treinamento

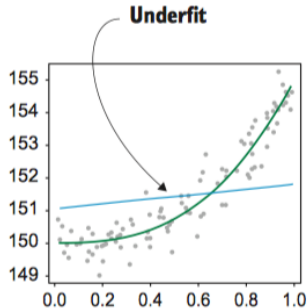
- Modelo aprende erros e ruídos
- Perde a capacidade de generalizar bem

- Pode ser evitado com técnicas de validação
- Ajuste de parâmetros

Overfitting







Sumarizando

- Aprendizado Supervisionado
- Problema de classificação

- Dados reais do problema
- Treinamento
- Teste
- Métricas de desempenho
- Comparar modelos
- Overfitting

Sumarizando