

Escola Superior de Tecnologia
Universidade do Estado do Amazonas

3 de novembro de 2018

Machine Learning

Hands on com Python

Samsung Ocean – Novembro/2018

Elloá B. Guedes da Costa

www.elloaguedes.com
ebgcosta@uea.edu.br

Machine Learning – Hands On com Python

Outline

- 1 Apresentação

Machine Learning – Hands On com Python

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



Machine Learning – Hands On com Python

Material do minicurso

<https://github.com/elloa/ocean-machineLearning2018>

<http://bit.ly/mlpython>

Machine Learning – Hands On com Python

Outline

2 Motivação

Machine Learning – Hands On com Python

Motivação



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 – Hands On com Python

Outline

3 Machine Learning I

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 – Hands On com Python

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

Machine Learning – Hands On com Python

Outline

- 4 Problema: Consumo de combustível

Problema Prático

- Problema: Consumo de combustível
- 1983, *American Statistical Association Exposition*
- Diversos modelos de carros
- Dados coletados de veículos de verdade
- Características particulares

Problema Prático

- Problema: Consumo de combustível
- 1983, *American Statistical Association Exposition*
- Diversos modelos de carros
- Dados coletados de veículos de verdade
- Características particulares
- Pergunta: Quantas milhas um dado carro faz com um galão de combustível?

Problema Prático

- Conjunto de dados: formato csv
- *Comma separated values*

Problema Prático

- Conjunto de dados: formato csv
- *Comma separated values*
- Valores separados por vírgulas
- Dataset multivariado
- Valores numéricos e nominais
- Dados faltantes

Problema Prático

- Vamos responder algumas perguntas iniciais a partir dos dados!
- Respostas com programação

Problema Prático

- Vamos responder algumas perguntas iniciais a partir dos dados!
- Respostas com programação
- Quantos exemplos de carros há no conjunto de dados?

Problema Prático

- Vamos responder algumas perguntas iniciais a partir dos dados!
- Respostas com programação
- Quantos exemplos de carros há no conjunto de dados?
- Quais são os atributos existentes no conjunto de dados?

Problema Prático

- Vamos responder algumas perguntas iniciais a partir dos dados!
- Respostas com programação
- Quantos exemplos de carros há no conjunto de dados?
- Quais são os atributos existentes no conjunto de dados?
- Quais os nomes dos carros existentes neste dataset?

Problema Prático

- Vamos responder algumas perguntas iniciais a partir dos dados!
- Respostas com programação
- Quantos exemplos de carros há no conjunto de dados?
- Quais são os atributos existentes no conjunto de dados?
- Quais os nomes dos carros existentes neste dataset?
- Quais as características do "chevrolet camaro"?

Problema Prático

- Vamos responder algumas perguntas iniciais a partir dos dados!
- Respostas com programação
- Quantos exemplos de carros há no conjunto de dados?
- Quais são os atributos existentes no conjunto de dados?
- Quais os nomes dos carros existentes neste dataset?
- Quais as características do "chevrolet camaro"?
- Qual a média de consumo dos carros existentes no dataset?

Problema Prático

- Vamos agora efetuar alguns ajustes nos dados!

Problema Prático

- Vamos agora efetuar alguns ajustes nos dados!
- Eliminar exemplos com dados faltantes

Problema Prático

- Vamos agora efetuar alguns ajustes nos dados!
- Eliminar exemplos com dados faltantes
- Eliminar coluna com nomes dos carros

Problema Prático

- Vamos agora efetuar alguns ajustes nos dados!
- Eliminar exemplos com dados faltantes
- Eliminar coluna com nomes dos carros
- Converter mpg para km/l
- $1 \text{ mpg} = 0.425 \text{ km/l}$
- Deletar coluna mpg

Machine Learning – Hands On com Python

Outline

5 Machine Learning II

Machine Learning

Processo de Aprendizagem de Máquina

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

Machine Learning – Hands On com Python

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

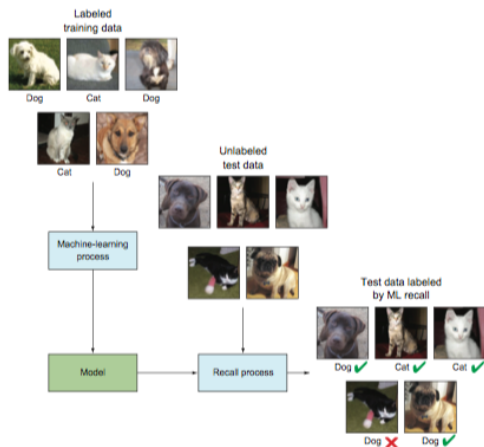
- Abordagem algorítmica
 - Programar todos os detalhes de como diferenciar gatos e cachorros
 - Texturas, cores, formas geométricas
 - Quantidade inviável de regras a serem capturadas
 - E se você esqueceu os gatos Sphynx ou os cachorros Komondor?
- Humanos cometem cerca de 7% de erros nesta tarefa

Machine Learning

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

Machine Learning – Hands On com Python

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 – Hands On com Python

Outline

6 Escolha do Algoritmo

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

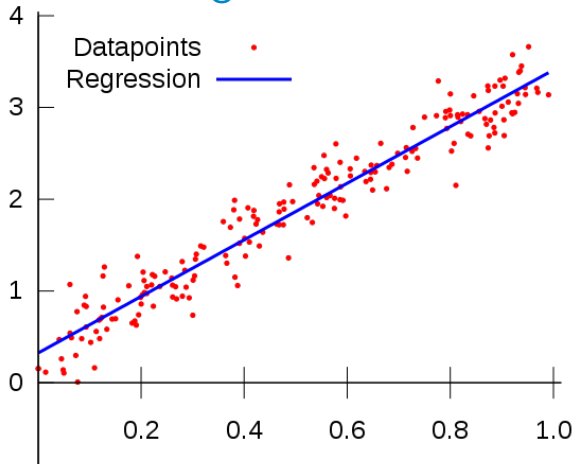
Escolha do Algoritmo

- 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 **regressão linear**
- **Ideia geral:** Assume que há uma relação linear entre os atributos preditores (X) e o atributo-alvo (Y)
- Encontra a reta que minimiza o erro

Escolha do Algoritmo



Escolha do Algoritmo

Criação do Modelo

```
from sklearn.linear_model import LinearRegression  
(...) regr = LinearRegression()
```


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

```
regr.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

- Vamos **visualizar** os resultados
- Cálculo dos resíduos
- Diferença ao quadrado entre o que foi previsto pelo modelo e o resultado do gabarito

Teste do Modelo

- Vamos **visualizar** os resultados
- Cálculo dos resíduos
- Diferença ao quadrado entre o que foi previsto pelo modelo e o resultado do gabarito
- O que se espera, em termos de resíduos, de um bom modelo?

Teste do Modelo

- Métricas de performance
- Problema de regressão
- Raiz do erro médio quadrático
- R^2

Teste do Modelo

Obtendo métricas de desempenho

```
from sklearn.metrics import mean_squared_error, r2_score  
(...)  
rmse = mean_squared_error(Y_predito, Y_test)  
r2 = r2_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 KNeighborsRegressor
(...)
# Inicializacao do modelo
kviz = KNeighborsRegressor(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?
- Algumas considerações:
 - O problema autmpg possui características fortemente lineares
 - Precisamos ter cuidado para prevenir **overfitting**

Machine Learning – Hands On com Python

Outline

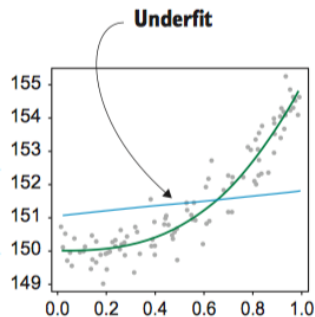
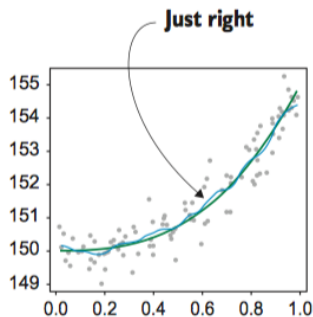
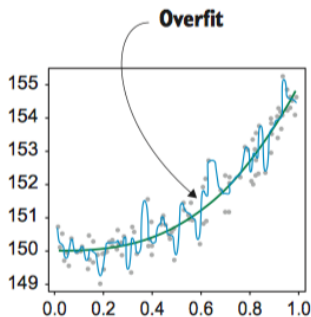
7 Overfitting

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

Machine Learning – Hands On com Python

Overfitting



Machine Learning – Hands On com Python

Outline

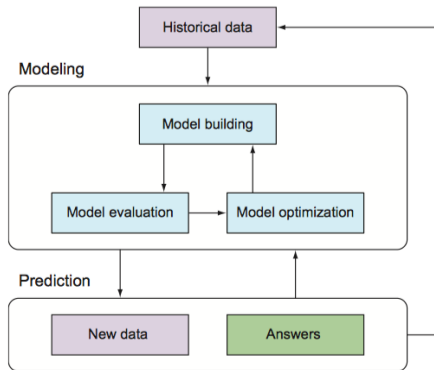
8 Sumarizando

Sumarizando

- Aprendizado Supervisionado
- Problema de regressão
- Dados reais do problema
- Treinamento
- Teste
- Métricas de desempenho
- Comparar modelos
- Overfitting

Machine Learning – Hands On com Python

Sumarizando



Machine Learning – Hands On com Python

Outline

9 Problema 2: Flores Iris

Problema 2: Iris Dataset



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

Problema 2: Iris Dataset



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

Problema 2: Iris Dataset

- Dadas as quatro medidas, qual a flor correspondente?

Problema 2: Iris Dataset

- Dadas as quatro medidas, qual a flor correspondente?
- Problema de classificação multi-classe
- Aprendizado supervisionado

Problema 2: Iris Dataset

- Dadas as quatro medidas, qual a flor correspondente?
- Problema de classificação multi-classe
- Aprendizado supervisionado
- Abra o dataset
- Atributos preditores e atributo alvo
- Dados já estão pré-processados
- Faça as partições de treinamento e teste (60/40)

Problema 2: Iris Dataset

- Vamos agora escolher um algoritmo para esta tarefa
- Árvores de decisão
- Hierarquia de perguntas
- Nós de divisão (perguntas)
- Nós de classificação (folhas)

Problema 2: Iris Dataset

Criando e treinando o modelo

```
from sklearn import tree
(...)
clf = tree.DecisionTreeClassifier()
clf.fit(X_train, Y_train)
```

Problema 2: Iris Dataset

- Vamos visualizar e entender a árvore gerada!
- Código está disponível nos notebooks de exemplo
- Abrir arquivo no github

Problema 2: Iris Dataset

Testando o modelo

```
resultados = clf.predict(X_test)
```

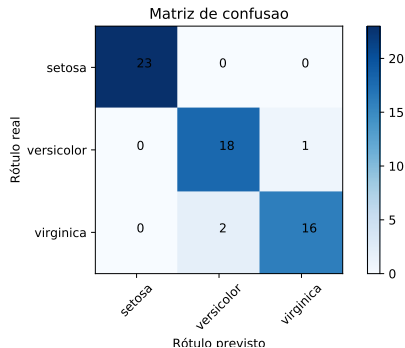
Problema 2: Iris Dataset

Matriz de confusão

```
from sklearn.metrics import confusion_matrix  
(...)  
confusion_matrix(Y_test, results)
```

Problema 2: Iris Dataset

- Vamos entender a matriz de confusão produzida



Machine Learning – Hands On com Python

Outline

- 10 Vantagens de Machine Learning

Vantagens de Machine Learning

- Acurado
 - Avaliação de métricas de desempenho
 - Auxílio no processo de tomada de decisão dos melhores parâmetros
 - Com mais dados, melhores ajustes podem ser feitos

Vantagens de Machine Learning

- Acurado
- Automatizado
 - Realizado com o auxílio do computador
 - Parâmetros e algoritmos ajustados automaticamente com os dados

Vantagens de Machine Learning

- Acurado
- Automatizado
- Rápido
 - Geração de resultados em questão de milisegundos, para a maioria dos problemas
 - Treinamento, teste e utilização

Vantagens de Machine Learning

- Acurado
- Automatizado
- Rápido
- Customizado
 - Considera os dados de cada domínio
 - Parâmetros podem ser ajustados para refletir métricas do negócio

Vantagens de Machine Learning

- Acurado
- Automatizado
- Rápido
- Customizado
- Escalável
 - Grande quantidade de dados
 - Computação em nuvem

Escola Superior de Tecnologia
Universidade do Estado do Amazonas

3 de novembro de 2018

Machine Learning

Hands on com Python

Samsung Ocean – Novembro/2018

Elloá B. Guedes da Costa

www.elloaguedes.com
ebgcosta@uea.edu.br