Big Data e Machine Learning com Hadoop e Spark

Conteúdo

CONTEÚDO PROGRAMÁTICO

- Visão geral da ciência de dados e aprendizado de máquina em escala
- Visão geral do ecossistema do Hadoop
- Instalação de um Cluster Hadoop
- Trabalhando com dados do HDFS e tabelas do Hive usando o Hue
- Visão geral do Python
- Visão geral do R
- Visão geral do Apache Spark 2
- Leitura e gravação de dados
- Inspeção da qualidade dos dados
- Limpeza e transformação de dados
- Resumindo e agrupando dados
- Combinando, dividindo e remodelando dados
- Explorando dados
- Configuração, monitoramento e solução de problemas de aplicativos Spark
- Visão geral do aprendizado de máquina no Spark MLlib
- Extraindo, transformando e selecionando recursos
- Construindo e avaliando modelos de regressão
- Construindo e avaliando modelos de classificação
- Construindo e avaliando modelos de cluster
- Validação cruzada de modelos e hiperparâmetros de ajuste
- Construção de pipelines de aprendizado de máquina
- Implantando modelos de aprendizado de máquina

MATERIAL DIDÁTICO

- Slides do treinamento em PDF
- GitHub com exercícios e códigos exemplo
- Máquinas virtuais para exercícios simulados
- Gravação das aulas disponível durante 3 meses

Teorema de Bayes

Teorema de Bayes

• Agora que entendemos a probabilidade condicional, podemos entender o Teorema de Bayes:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Descrição - a probabilidade de A dado B, é a probabilidade de A vezes o probabilidade de B dado A sobre a probabilidade de B.
O principal insight é que a probabilidade de algo que depende de B depende muito sobre a probabilidade básica de B e A. As pessoas ignoram isso o tempo todo.

Caso do Teorema de Bayes

• O teste de drogas é um exemplo comum. Mesmo um “altamente preciso” teste de drogas pode produzir mais falso-positivos do que verdadeiro-positivos.
• Digamos que tenhamos um teste de drogas que possa determinar com precisão identificar usuários de uma droga 99% do tempo, e com precisão tem um resultado negativo para 99% de não usuários. Mas apenas 0,3% da população total realmente usa essa droga.

Teorema de Bayes

• Evento A = É um usuário do medicamento, Evento B = testado positivamente para o medicamento.
• Podemos calcular a partir dessa informação que P (B) é de 1,3% (0,99 * 0,003 + 0,01 * 0,997 - a probabilidade de teste positivo se você usar, mais o probabilidade de teste positivo se você não fizer isso.

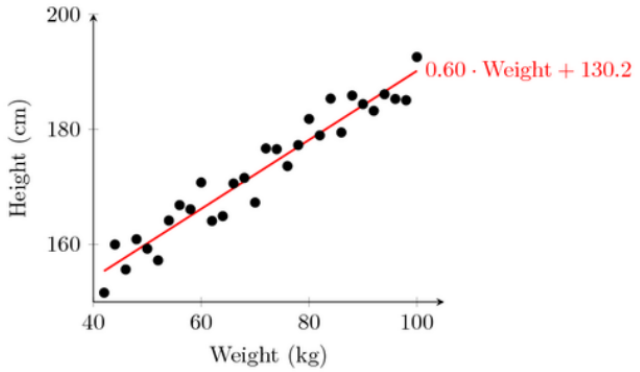
$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} = \frac{0.003 * 0.99}{0.013} = 22.8\%$$

• Então, as chances de alguém ser um usuário real da droga, dado que eles testado positivo é de apenas 22,8%!
• Embora P (B | A) seja alto (99%), não significa que P (A | B) esteja alto.

Regressão Linear

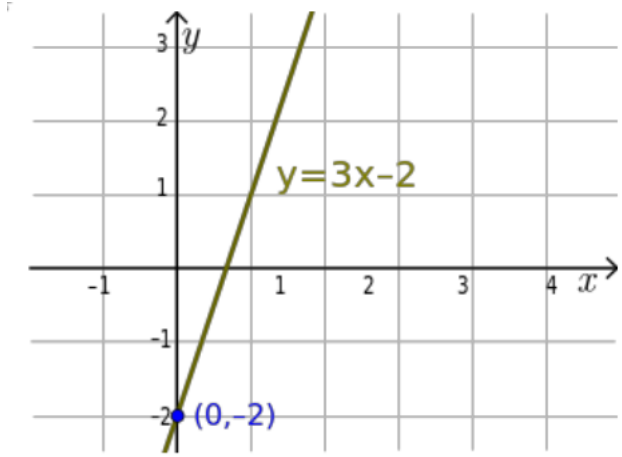
Regressão Linear

• Ajustar uma linha a um conjunto de dados de observações
• Use esta linha para prever valores não observados
• Eu não sei por que eles chamam de "regressão".É realmente enganador. Você pode usá-lo para prever pontos no futuro, o passado, tanto faz. Na verdade, o tempo geralmente não tem nada a ver com isso.



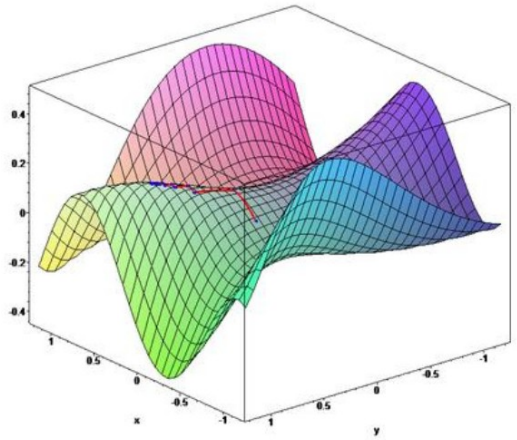
Regressão Linear: como funciona?

• “Mínimos quadrados” minimiza a soma dos erros quadrados.
• Isto é o mesmo que maximizar a probabilidade dos dados observados se você começar a pensar no problema em termos de probabilidades e probabilidades funções de distribuição
• Isso às vezes é chamado de “estimativa de máxima verossimilhança”



Mais de uma maneira de fazer isso

- Gradiente descendente é um método alternativo aos mínimos quadrados.
- Basicamente itera para encontrar a linha que melhor segue os contornos definidos pelos dados.
- Pode fazer sentido quando se lida com dados 3D
- Fácil de experimentar em Python e apenas comparar resultados para mínimos quadrados
  - Mas geralmente os mínimos quadrados são perfeitamente boas escolha.



Medição de Erro com R-Quadrado

- Como medimos quão bem nossa linha se ajusta aos nossos dados?
- Medidas de R-Quadrado (coeficiente de determinação):

A fração da variação total em Y que é capturado pelo modelo

Computação R-Quadrado

$$1{,}0 - \frac{\text{soma de erros quadrados}}{\text{soma da variação quadrática da média}}$$

Interpretando o R-Quadrado

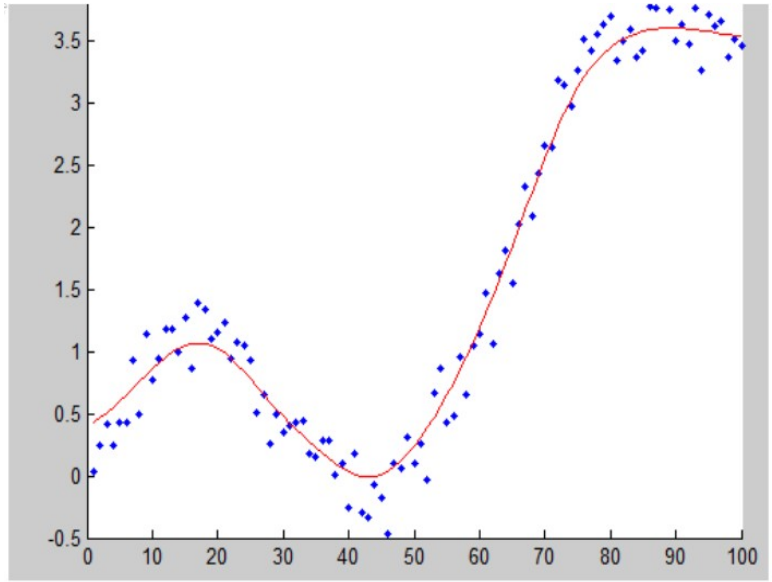
- Varia de 0 a 1
- 0 é ruim (nenhuma das variações é capturada), 1 é bom (todas as variações são capturadas).

Vamos ver um exemplo

Regressão Polinomial

Por que nos limitar a linhas retas?

- Nem todos os relacionamentos são lineares.
- Fórmula linear: y = mx + b
  - Este é de "primeira ordem" ou "primeiro grau" polinomial, com a potência de x = 1
- Polinômio de segunda ordem: y = ax 2 + bx + c
- Terceira ordem: y = ax 3 + bx 2 + cx + d
- Ordens mais altas produzem curvas mais complexas.



Cuidado com o overfitting

• Não use mais graus do que você precisa
• Visualize seus dados primeiro para ver quão complexa pode realmente ser uma curva
• Visualize o ajuste - sua curva está saindo de seu caminho para acomodar outliers?
• Um r-quadrado elevado significa simplesmente que sua curva se ajusta bem aos seus *dados de treinamento*; mas pode não ser um bom preditor.
• Mais tarde, falaremos sobre maneiras mais fundamentadas de detectar overfitting
(treinamento / teste)

Exemplo

Regressão Multivariada

Regressão Multivariada (Regressão Múltipla)

- E se mais de uma variável influencia no que você está interessado?
- Exemplo: prevendo um preço para um carro baseado em seus muitos atributos (estilo da carroceria, marca, quilometragem, etc.)

Ainda usa menor-quadrado

- Acabamos com coeficientes para cada fator.
  - Por exemplo, preço = $\alpha + \beta_1$ quilometragem + β_2 idade + β_3 portas
  - Esses coeficientes indicam quão importante é cada fator (se os dados são todos normalizados!)
  - Livre-se daqueles que não importam!
- Ainda pode medir o ajuste com o r-quadrado
- É necessário assumir que os diferentes fatores não são dependentes de entre si.

Exemplo

Modelos de Múltiplos Níveis

Modelos de Múltiplos Níveis

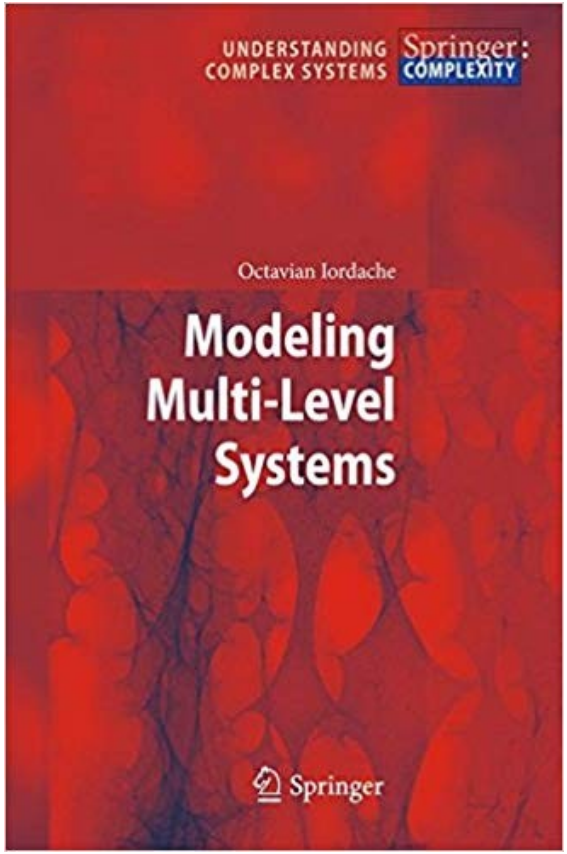
• O conceito é que alguns efeitos acontecem em vários níveis.
• Exemplo: sua saúde depende de uma hierarquia da saúde de suas células,
órgãos, você como um todo, sua família, sua cidade e o mundo em que vive.
• Sua riqueza depende do seu próprio trabalho, do que seus pais fizeram, do que
seus avós fizeram, etc.
• Modelos multiníveis tentam modelar e contabilizar essas interdependências.

Modelando vários níveis

• Você deve identificar os fatores que afetam o resultado que você está tentando prever em cada nível.
• Por exemplo - os resultados do SAT podem ser previstos com base na genética de
crianças individuais, o ambiente doméstico de crianças individuais, o taxa de criminalidade do bairro em que vivem, a qualidade dos professores em sua escola, o financiamento de seu distrito escolar, e a educação políticas do seu estado.
• Alguns desses fatores afetam mais de um nível. Por exemplo, crime taxa pode influenciar o ambiente doméstico também.

Fazer isso é difícil.

- Esteja ciente do conceito, uma vez que modelos multi-nível aparecem em alguns requisitos de trabalho de ciência de dados.
- Tópico mais avaçado.



Aprendizado de Máquina Supervisionado e Não Supervisionado

E o conceito de treinamento / teste

O que é aprendizado de máquina?

• Algoritmos que podem aprender com dados observacionais, e pode fazer previsões baseadas neles.

Aprendizagem não supervisionada

• O modelo não recebe “respostas” para aprender; deve entender os dados apenas pelas próprias observações.
• Exemplo: agrupe alguns objetos em dois conjuntos diferentes. Mas eu não digo qual é o conjunto "certo" para qualquer objeto antes do tempo.




Eu quero coisas grandes e pequenas? Coisas redondas e quadradas? Coisas vermelhas e azuis?
O aprendizado não supervisionado poderia me dar qualquer um desses resultados.

Aprendizagem não supervisionada

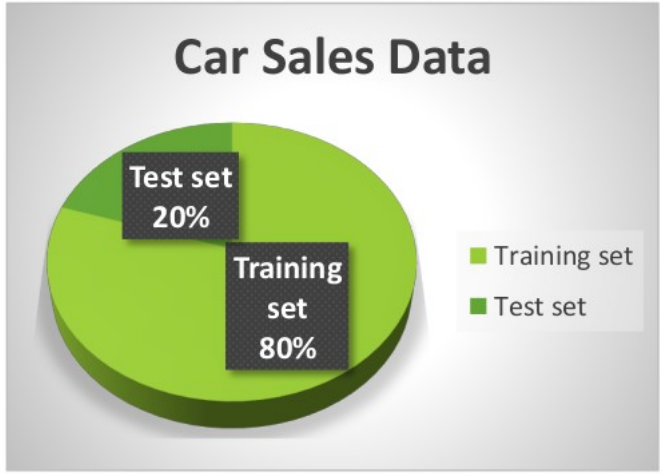
• Aprendizado não supervisionado soa horrível! Por que usar isso?
• Talvez você não saiba o que está procurando - você está procurando variáveis latentes.
Exemplo: agrupar usuários em um site de namoro com base em suas informações e comportamento. Talvez você encontre grupos de pessoas que surgem que não se combinam aos seus estereótipos conhecidos.
• Grupos de filmes com base em suas propriedades. Talvez nossos conceitos atuais de gênero estão desatualizados?
• Analise o texto das descrições dos produtos para encontrar os termos que carregam mais significado para uma determinada categoria.

Aprendizagem Supervisionada

- Na aprendizagem supervisionada, os dados que o algoritmo “aprende” vem com as respostas “corretas”.
- O modelo criado é então usado para prever a resposta para novos valores desconhecidos.
- Exemplo: você pode treinar um modelo para prever os preços de carros baseados em atributos de carros usando dados de vendas históricas dados. Esse modelo pode então prever o preço ótimo para carros novos que não foram vendidos antes.

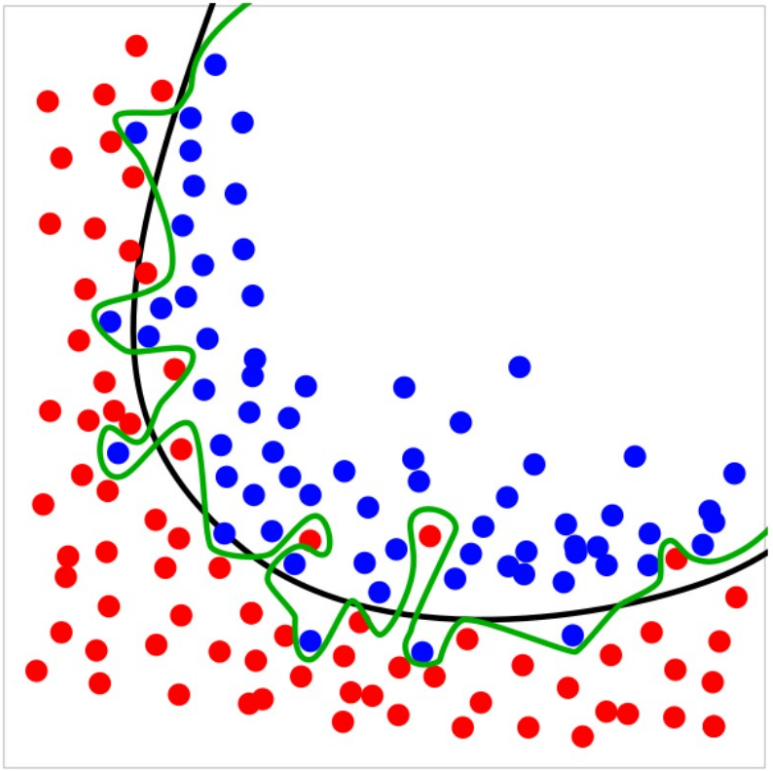
Avaliando o Aprendizado Supervisionado

• Se você tiver um conjunto de dados de treinamento que inclua valor que você está tentando prever - você não precisa adivinhar se o modelo resultante é bom ou não.
• Se você tiver dados de treinamento suficientes, poderá dividi-lo em duas partes: um conjunto de *treinamento* e um conjunto de *teste*.
• Você então treina o modelo usando apenas o conjunto de treinamento
• E, em seguida, medir (usando r-quadrado ou alguma outra métrica) a precisão do modelo, pedindo-lhe para prever valores para o conjunto de teste e compare isso com o valores conhecidos e verdadeiros.



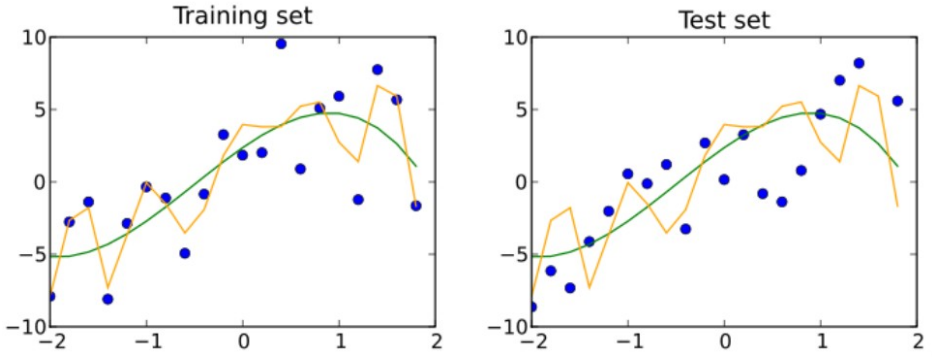
Treinar / Testar na prática

- Necessidade de garantir que ambos os conjuntos sejam grandes o suficiente para conter representantes de todas as variações e outliers nos dados
- Os conjuntos de dados devem ser selecionados aleatoriamente
- Treinar / testar é uma ótima maneira de proteger contra overfitting



Treinamento / teste não é infalível

- Talvez seus tamanhos de amostra sejam muito pequenos
- Ou devido à chance aleatória de seu conjuntos de treino e de teste parecerem notavelmente semelhantes
- Overfitting ainda pode acontecer



Validação Cruzada K-fold

- Uma maneira de proteger ainda mais o overfitting é a validação cruzada K-fold
- Parece complicado. Mas é uma ideia simples:
  - Divida seus dados em K segmentos atribuídos aleatoriamente
  - Reserve um segmento como seus dados de teste
  - Treine em cada um dos segmentos restantes do K-1 e meça seus desempenhos contra o conjunto de teste
  - Pegue a média dos k-1 escores r-quadrado

Exemplo

Obrigado!!!

Nos vemos amanhã!!!

Bom descanso!