Modelagem Analítica com Machine Learning

Primeiro Dia

Paulo Cysne Rios, Jr.

Objetivos

- Entender as principais modelagens analíticas de dados estruturados atuais.
- Conhecer os seus conceitos, seus princípios, as suas vantagens e desvantagens.
- Saber quando usar que modelagem.
- E quando não usar que modelagem.

Objetivos

- Conhecer vários cases com estas modelagens.

- Aplicar modelagem analítica a série temporais.

Objetivos

- Ao final do curso, você estará em condições de fazer modelos analíticos de dados estruturados.

- Dados estruturados = aqueles que podem ser considerados em forma de tabelas.

Observação

- Aprenderemos sobre modelagens de dados não estruturados (imagem, audio, etc) no curso de Introdução a Deep Learning.

- Lá veremos também como a modelagem de Deep Learning pode ser aplicada a séries temporais.

- E a alguns tipos de dados estruturados também!

- Mas o conhecimento deste curso é fundamental para saber quando usar qual modelagem, inclusive Deep Learning!

Pré-Requisitos

- Conhecimentos de Python, Numpy e Pandas.
- Conhecimento da identificação dos objetivos de projetos de analítica de dados.
- Conhecimentos do pré-processamento usado em projetos de analítica de dados.
- Conhecimento das validações, testes e otimizações feitos em projetos de analítica de dados.
- Ideal: conhecimento de Álgebra Linear e Cálculo Diferencial.

Formação em ML

- Introdução a ML - obter os conhecimentos básicos (Numpy, Pandas, pré-processamento de dados, estrutura de um projeto de analítica de dados, validação e testes de modelos, cases).

- Modelagem Analítica em ML - fazer modelos analíticos em dados estruturados e em séries temporais.

- Introdução a Deep Learning - fazer modelos analíticos em séries temporais, dados não estruturados e alguns tipos de dados estruturados.

- Visualização de Dados em ML - usar a visualização de dados para comunicação com outros, para exploração dos dados e para pré-processamento de dados.

Revisão Rápida

Objetivo de Um Projeto de Analítica de Dados

- Ele deve ser estratégico.
- Ele deve usar conjunto de dados existentes para explorar e identificar padrões, tendências e relacionamentos neles existentes.
- Começa sempre pela pergunta: o que desejamos fazer?
- Se você não tiver um objetivo claro, não vai saber aonde chegar!
- Qual é a natureza preditiva do seu objetivo?

Pré-Processamento

- Os dados devem estar de uma forma que podem ser usados para a modelagem.

- Na grande maioria das vezes eles não estão!

- O que você deve fazer com os dados também depende do tipo de modelagem e dos objetivos de seu projeto.

Pré-Processamento

- Importar e ler os dados de várias fontes ou de uma fonte.
- Identificar e lidar com valores que faltam (valores nulos): remover, substituir pela média total ou pela média de uma classe.
- Identificar e lidar com outliers: remover, ajustar, corrigir, falar com pessoas da área.

Pré-Processamento

- Identificar e lidar com valores não inválidos (por exemplo, idade = -4).

- Identificar e lidar com escalas diferentes (por exemplo, uma variável vai de 0 a 1, outra de vai de 10 a 1000). Alguns modelos analíticos exigem que todos dados estejam em escalas semelhantes.

- Identificar e codificar numericamente valores com categorias ou texto.

Pré-Processamento

- Identificar a distribuição dos dados, se ela é uma distribuição normal (formato de sino) ou não.
- Identificar correlações entre as variáveis.
- Identificar e lidar com valores que faltam (valores nulos): remover, substituir pela média total ou pela média de uma class.
- Dividir os conjuntos de dados em treinamento e teste.

Medidas de Desempenho da Modelagem

- Classificação = Tabela de Confusão, Acurácia, Precisão, Recall, Gráfico de Curva ROC, Precisão e Recall Tradeoff.

- Regressão = RMSE (Root Mean Square Error), score r2.

Validação e Teste

- Validar = verificar no conjunto de dados de treinamento o modelo que foi treinado nele (ou numa parte dele).

- Usando Validação Cruzada, se treina o modelo numa parte do conjunto de dados de treinamento e se valida noutra parte deste, aleatoriamente.

- Testar = verificar no conjunto de dados de teste o que modelo que foi treinado no conjunto de dados de treinamento.

Overfitting e Underfitting

- Overfitting = Os resultados da verificação são bem melhores que os resultados do teste.

- Boa modelagem = Os resultados da verificação e do teste são bons. Quão bom depende dos objetivos do projeto em questão!

- Underfitting = Os resultados da verificação não são bons. Os do teste também não.

Otimização

- Uma vez que uma modelagem analítica foi escolhida, se procura os melhores valores de seus hiperparâmetros (parâmetros do modelo).
- Para isso se usa uma técnica conhecida como Grid Search.
- Nela se mede o desempenho do modelo para diferentes combinações dos valores de seus hiperparâmetros.

Tipos de Aprendizagem

- Aprendizagem Supervisionada:
- Existem dados históricos com o valor objetivo, os dados históricos tem um label, o que se busca nos novos valores.
- Aprendizagem Não Supervisionada:
- Não existem dados com o valor objetivo. O que se procura é encontrar grupos ou uma estrutura.

Aprendizagem Supervisionada

- Tipo Regressivo:

- O valor objetivo é um valor numérico contínuo.

- Tipo Classificação:

- O valor objetivo é um valor de uma categoria ou um número discreto, não contínuo. O valor objetivo pertence a uma classe.

Aprendizagem Supervisionada

Notação

- y é o valor objetivo real, como está nos dados, em forma de um vetor.

- ŷ é o valor objetivo predito pelo modelo analítico, em forma de um vetor.

- y é a variável dependente, o label, o objetivo.

Notação

- X é uma matriz com os valores das variáveis independentes, aquelas usados para predizer o valor y.
- X tem m linhas = número de observações.
- X tem n colunas = número de variáveis independentes.
- Para cada linha de X há um valor y correspondente e um $\hat{y}$ correspondente.

Notação

- $x^{(i)}$ é um vetor com todos os valores das variáveis independentes da linha/observação i. Ele não inclui o valor buscado y.

- $y^{(i)}$ é o valor objetivo, buscado para a linha/observação i.

Exemplo

Conjunto de dados com os valores médios de imóveis nos distritos de Boston

$$\mathbf{x}^{(1)} = \begin{pmatrix} -118.29 \\ 33.91 \\ 1,416 \\ 38,372 \end{pmatrix}$$

Longitude

Latitude

Número de habitantes

Salário anual em dólares

$$y^{(1)} = 156,400$$

Valor médio da casa

Exemplo

A matriz X neste caso

$$\mathbf{X} = \begin{pmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \vdots \\ (\mathbf{x}^{(1999)})^T \\ (\mathbf{x}^{(2000)})^T \end{pmatrix} = \begin{pmatrix} -118.29 & 33.91 & 1,416 & 38,372 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

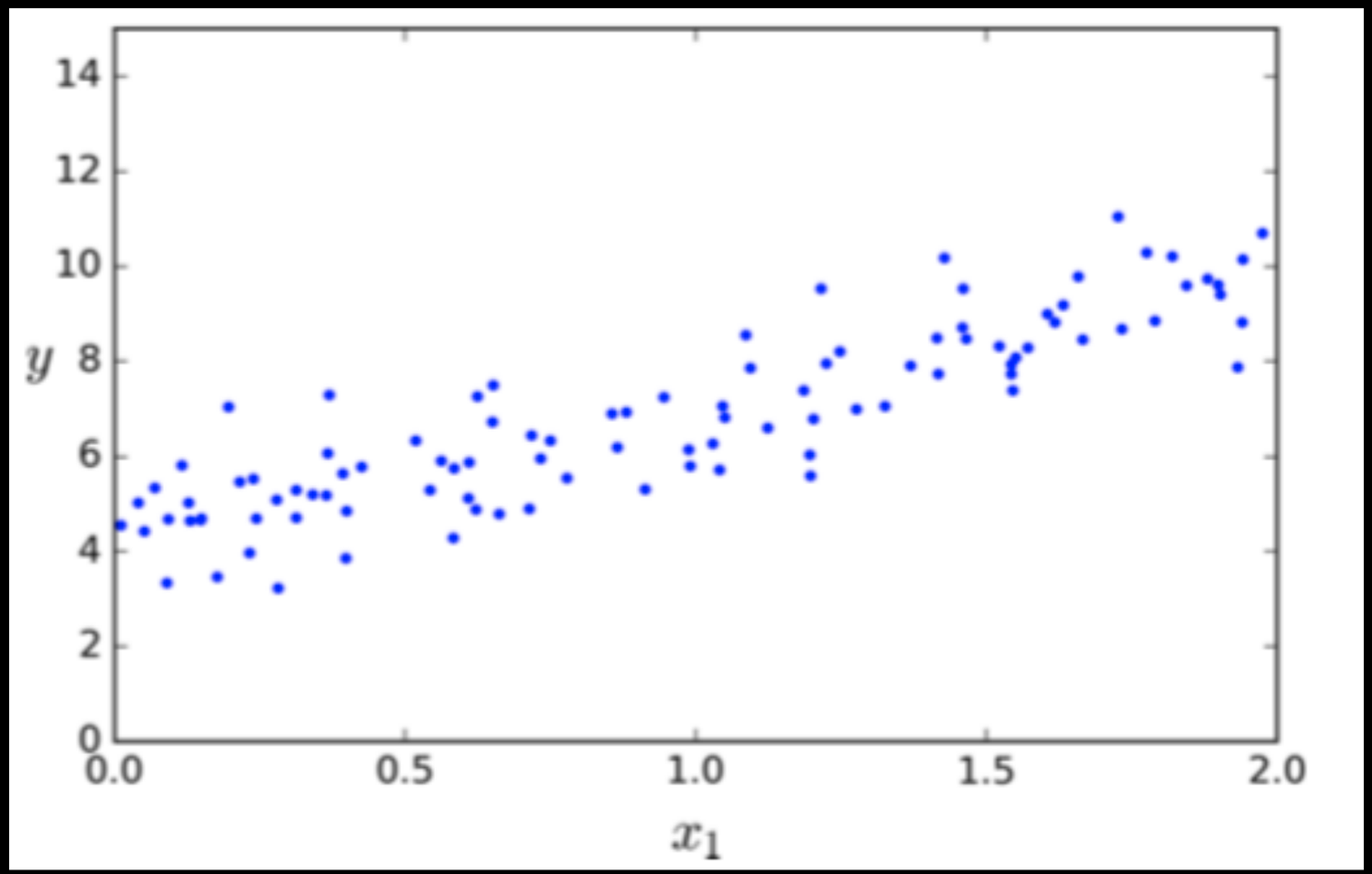
Note que X^T é a matriz transposta de X

Regressão Linear

Quando se usa Regressão Linear

- Há uma relação diretamente ou indiretamente proporcional entre as variáveis independentes e a variável dependente.
- Esta relação é da forma y = a + bx.
- Quando somente há uma variável independente.

Quando se usa Regressão Linear



Mais de uma variável independente

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

De forma matricial

$$\hat{y} = h_\theta(\mathbf{x}) = \theta^T \cdot \mathbf{x}$$

Diferença

A diferença entre o valor predito e o valor real
pode ser expresso através do mean square error (MSE)

$$\mathrm{MSE}(\mathbf{X}, h_\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(\theta^T \cdot \mathbf{x}^{(i)} - y^{(i)}\right)^2$$

Nosso objetivo é minimizar este valor!!

Minimizando o custo

Esta diferença é a nossa função custo

Se pode matematicamente minimizar esta função custo

Escolhendo coeficientes que a minimizem

Matematicamente se pode provar que estes valores dos coeficientes minimizam esta função custo

$$\hat{\theta} = \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$$

Exemplo

no Jupyter Notebook

Exercício Prático

Crie o seguinte exemplo

- Use o Jupyter notebook.
- Crie um case de conjunto de dados com valores aleatórios, fazendo uma função real y = a + xb. Seja criativo. Por exemplo, x é o BMI da pessoa (de 17 a 50), y é sua glicose (de 50 a 800),
- Faça uma modelagem de regressão linear usando a equação de mínimo custo e usando Scikit-Learn.
- Plote o modelo com os dados reais.
- Compare as 2 soluções: com a equação e com Scikit-Learn.

Consequências da Equação de Minimização do Custo

$$\hat{\theta} = \left(\mathbf{X}^T \cdot \mathbf{X}\right)^{-1} \cdot \mathbf{X}^T \cdot \mathbf{y}$$

- A equação depende do produto de matrizes X.

- Sua computação se torna bastante lenta quando o número de variáveis independentes se torna muito grande (por exemplo, acima de 100 mil como no caso de aplicações em biologia molecular).

Consequências da Equação de Minimização do Custo

- Mas esta equação é linear em relação ao número de observações/linhas.

- Quer dizer, ela pode lidar com conjuntos de dados bastante grandes (muitas instâncias/observações).

- Conquanto que estes encontrem lugar na memória do computador (assumindo não se ter nenhuma estrutura de processamento distribuído como Hadoop/Stark).