ESPECIALIZAÇÃO EM
MACHINE LEARNING

Regressão linear com folhas de cálculo

Módulo 1 – Tema 1

Tokio.

Instruções

Vamos **simular um modelo matemático**, com dados fictícios, para compreender melhor como realizar uma regressão linear usando, neste caso, um software de folhas de cálculo, como LibreOffice Calc, Microsoft Excel, Google Sheets, etc.

Para isso, primeiro escolhemos um destes softwares, com o qual estejamos mais à vontade, utilizando as suas diferentes secções.

Se necessário, podemos descarregar qualquer um deles. Especialmente, o LibreOffice é um pacote de burótica gratuito, o Google Sheets é gratuito para todos os utilizadores do GMail.

Como cada formando pode utilizar um software diferente, não podemos explicar os passos para cada um deles, mas sim no geral. Se ocorrer algum problema, podemos procurar rapidamente a solução através de qualquer motor de busca, ou envie um email ao formador pela plataforma.

ESPECIALIZAÇÃO EM MACHINE LEARNING

Na plataforma, existem vários arquivos com conjuntos de dados diferentes. Segundo a tarefa, utilizaremos o respetivo conjunto de dados.

Recomendamos a localização das colunas referenciadas em cada tarefa do conjunto de dados pedido das e importá-las-emos para um novo documento, uma vez que iremos trabalhar com diferentes colunas consoante o que é pedido na tarefa.

Podemos trabalhar com vários documentos, um por tarefa, ou melhor, com várias folhas no mesmo documento de cálculo.

Para carregar a prática poderemos fazê-lo através de um documento com várias folhas ou em documentos diferentes. Para enviar o exercício, **nomearemos os documentos, ou as folhas dentro de um único documento, de forma descritiva**, explicando que tarefa resolve cada um deles.

Tokio.

Exemplo: modelo e dados

Neste exemplo, vamos simular um modelo de econometria: uma curva de preço, com o preço de um produto vs. o seu número de vendas.

Os modelos matemáticos/estatísticos/científicos são habitualmente usados para:

- Explicar a relação entre 2 ou mais variáveis: a variável dependente do resto (a *Y*) e as variáveis independentes (a *X*), ou *modelar o seu comportamento*.
- A partir desse modelo, realizar predições no futuro, sejam *interpolações* ou *extrapolações*.

Iremos discutir ambos os objetivos durante a prática.

Modelo de regressão linear por mínimos quadrados

Vamos realizar uma **regressão linear por mínimos quadrados** utilizando os dados disponibilizados. Esta regressão linear modelará os dados, ou seja, criará um modelo matemático a partir dos mesmos.

O conjunto de dados, também chamado "dataset" em inglês, irá conter os nossos dados. Verificámos como afeta o preço de venda de um produto com o seu número de vendas, através de vários testes: diferentes variações do produto, descontos e campanhas promocionais, diferentes mercados, vendas privadas, etc.

O modelo matemático relacionará duas variáveis:

- A variável Y será o nº de vendas, a variável objetivo ou dependente.
- A variável X será o preço do produto, a variável independente.

Simplificámos os nomes em X e Y para facilitar o exercício.

Tarefa 1

Dados a usar: módulo1-unidade1-exercício1-dataset-tarefa1.csv

Passo 1

Antes de treinar um modelo, devemos sempre tentar visualizar os dados. Num modelo com um dataset muito mais completo e complexo, é habitualmente mais complicado, mas numa regressão linear simples, é bastante fácil.

Criamos um gráfico de dispersão (ou "scatter plot") e mostramos no eixo horizontal ou *X* a variável *X*, e no eixo vertical ou *Y*, a variável *Y*.

Numa célula do documento/folha, indicamos "**Pergunta 1**" e respondemos à pergunta: Qual a forma seguida pelo gráfico apresentado (a que tipo de função se enquadra) ? *Por que se trata, neste caso, de regressão **linear**?*

Tarefa 1

Passo 2

Vamos “modelar os dados” (como diríamos em estatística) ou “treinar o modelo” (como diríamos em machine learning).

Para isso, calculamos os valores de *m* e *b*, segundo as fórmulas seguintes :

$$Y = m \times X + b$$

$$m = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

$$b = \overline{y} - m \times \overline{x}$$

Tarefa 1

Passo 2

Notas:

- As variáveis, com uma barra horizontal superior, indicam a média da variável e o símbolo Σ indica o somatório, ou soma de todos os valores da variável.
- Indicar as células que irão conter os valores de m e b, de uma forma descritiva, face à avaliação da prática.
- Podemos usar as funções integradas do nosso software de folhas de cálculo para encontrar as médias e somatórios, ou fazê-lo de forma manual.
- Recomendamos a utilização de outras células auxiliares para calcular valores intermédios, quando for necessário.

Com estes coeficientes, definimos o nosso modelo. Estes coeficientes m e b são os que nos permitem explicar o comportamento das variáveis ou realizar predições com o modelo.

Tarefa 1

Passo 3

Vamos avaliar o modelo, utilizando o coeficiente de correlação.

Podemos calculá-lo com as fórmulas(Coeficiente de correlação = CoVariança(x,y) / Desvio Padrão(x) * Desvio Padrão(y)):

$$R^2 = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y};$$

$$\sigma_{xy} = \frac{\overline{x \cdot y}}{n};$$

$$\sigma_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2};$$

$$\sigma_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}$$

Tarefa 1

Passo 3

Notas:

- Indicamos numa célula, de forma clara, o valor de R^2.
- Para calcular $X \cdot Y$, X^2 ou Y^2, podemos criar colunas auxiliares, a partir das colunas originais, multiplicando os seus valores ou elevando-os ao quadrado.
- Podemos verificar os cálculos com as funções de desvio-padrão e covariância do software de folhas de cálculo (cuidado: deveremos usar a função de desvio-padrão para toda a população, não para uma amostra, podem ser funções diferentes).

Numa célula do documento/folha, indicamos "**Pergunta 2**" e respondemos à pergunta: *O que significa o valor de R^2?*

Tarefa 1

Passo 4

Vamos calcular os valores de *Y* que o modelo prediria para cada valor de *X*, segundo esses valores de *m* e *b*.

Para isso, criamos outra nova coluna chamada *y_pred* e calculamos esses valores, segundo a seguinte fórmula:

$$Y = m \times X + b$$

Em alguma célula do documento/folha, indicamos "**Pergunta 3**" e respondemos à pergunta: Qual a relação entre os resultados dessa coluna e o valor de R^2?

Passo 5

Este passo é opcional.

No software de folha de cálculo, voltamos ao gráfico e adicionamos uma linha de tendência e calculamos o seu R2, usando diretamente as funcionalidades do software, habitualmente disponíveis nos gráficos de dispersão.

Tarefa 2

Dados a usar: módulo1-unidade1-exercício1-dataset-tarefa2.csv

Devemos trabalhar numa folha ou documento diferente para esta tarefa, importando as colunas do dataset a utilizar em cada um dos passos.

Passo 1

Até agora temos usado dados simulados sem qualquer erro, com uma correlação direta perfeita, o que não acontece habitualmente na vida real.

Repetimos os passos, da tarefa anterior, do 1 ao 4 (criar o gráfico, calcular *m* e *b*, calcular R^2 e calcular *y_pred*), só não respondemos às perguntas, para os dados das colunas *X_real* e *Y_real*.

Incluímos a coluna *y_pred* como uma nova série no gráfico original de *X_real* e *Y_real*. Se for possível, (nem sempre é), incluímos esta série como um gráfico de linhas em vez de pontos, para que seja mais fácil de visualizar a linha de tendência.

Tarefa 2

Passo 2

Agora vamos calcular os *resíduos*. Os resíduos são a diferença entre *Y* real e *y_pred*. Iremos calculá-los para cada valor de *X* e criamos um novo gráfico, onde representemos os resíduos no eixo vertical e X no eixo horizontal.

Os residuos neste exercicio são valores aleatórios, que seguem uma distribuição normal, que podemos denominar *ruído*. Esses resíduos, podem corresponder a erros nas medidas das variáveis, diferenças aleatórias nas variaveris, variáveis ocultas que não temos em conta que deveriam entrar no calculo, etc.

Passo 3

Por outro lado, vamos utilizar este novo modelo, para realizar predições sobre novos valores de 2 tipos:

- A interpolação é realizar predições sobre valores na mesma classe que o dataset original, entre o valor máximo e mínimo.
- A extrapolação é realizar predições sobre valores fora da classe do dataset original, com valores inferiores ao mínimo ou superiores ao máximo.

Para isso, escolhemos 6 valores aleatórios, 3 deles da classe do dataset original e outros 3 fora da classe, e predizemos os seus valores de *y_pred*.

Tarefa 3

Dados a usar: módulo1-unidade1-exercício1-dataset-tarefa3.csv

Passo 1

Vamos repetir os passos da tarefa anterior num novo dataset, usando as colunas *X_erro* e *Y_erro*.

Para esses dados, criamos os gráficos, calculamos *m* e *b*, calculamos R^2, *y_pred*, adicionamos à linha de tendência *y_pred* ao gráfico e criamos o gráfico de resíduos.

Neste caso, analisando os resultados e os gráficos, e não só R^2, podemos ver que o nosso modelo é muito mais fraco que os anteriores.

Seríamos capazes de verificar o porque de o modelo não funciona tão bem, antecipadamente, só através da visualização do gráfico de dispersão com X_erro e Y_erro?

Tarefa 3

Passo 2

Analisamos o gráfico de *X_erro* e *Y_erro*:

Que relação diríamos que têm os dados originais? É uma relação linear ou de outro tipo?

Passo 3

Como poderíamos transformar os dados originais para os poder modelar, usando regressão linear simples?

Pistas:

- Criar uma nova coluna a partir da *X* dos dados originais.
- Analisar, em detalhe, o gráfico de dispersão de *X* vs. *Y*.
- Transformar os dados originais de alguma forma, por exemplo, somando-lhes um valor, multiplicando-os por um valor, elevando-os a um número, passando-os por alguma função, etc.
- *A resposta tem '4 lados iguais'. :)*

Tarefa 4

Dados a usar: módulo1-unidade1-exercício1-dataset-tarefa4.csv

Neste caso, carregamos os dados *X_rand* e *Y_rand e representamo-los num gráfico.*

Numa célula do documento/folha, indicamos "**Pergunta 4**" e respondemos à pergunta: Podemos treinar um modelo que encontre algum tipo de relação linear entre ambas as variáveis, inclusivamente transformando os dados? Porquê?

Tarefa 5

Nesta tarefa não vamos usar nenhum dataset, pois vamos cria-lo

Vamos simular dados sintéticos que sigam uma determinada relação, para poder gerar datasets de teste e verificar os nossos algoritmos e implementações de machine learning.

Para isso, geramos uma nova folha ou documento onde vamos criar várias colunas, segundo as instruções seguintes.

Passo 1

Vamos gerar um dataset semelhante ao que usámos na primeira tarefa. Para isso seguimos os passos seguintes:

- Criar uma coluna *X_passo1*, com valores na classe *[0, 10, 20, …, 100]*.
- Criar 2 células para os valores de *m* e *b* e dar-lhes 2 valores aleatórios.
- Gerar outra coluna calculada, *Y_passo1*, utilizando os valores de *m*, *b* e *X_passo1*.

Tarefa 5

Passo 2

Agora vamos gerar um dataset com ruído aleatório. Para isso, seguimos os passos do ponto anterior para gerar 2 colunas *X_passo2* e *Y_passo2*, e além disso:

- Vamos adicionar ruído à coluna *Y_passo2*. numa nova coluna.
- Criamos uma nova célula auxiliar, que representará a escala de erro. Pode estar em unidades, dezenas, centenas, decimais, ou até números negativos.
- Criamos uma nova coluna, *Y_passo2_erro*, com a fórmula:
  *Y_passo2_erro = Y_passo2 + Y_passo2 * N * e*.
- *N* é um número aleatório na classe [-1, 1], que podemos gerar com a fórmula, usando a função de geração de números aleatórios (normais ou não) do software de folha de cálculo:
  *(RANDN - 0,5) * 2*
- Representamos num gráfico de dispersão os valores das colunas *X_passo2, Y_passo2* e *Y_passo2_erro*.
- Alteramos o valor de *e*, vendo como afeta os valores de *Y_passo2_erro*, até que fique um erro mais ou menos *normal*.

Tarefa 5

Passo 3

Neste passo, vamos gerar um dataset com dados polinómios. Para isso, seguiremos os passos do passo 1 para gerar 2 colunas *X_passo3* e *Y_passo3*, e além disso:

- Podemos usar uma nova folha ou documento se desejarmos.
- Criamos uma nova coluna, *X_passo3_quadrado*, com os valores de *X_passo3* elevados ao quadrado.
- Mudamos a coluna *Y_passo3* para que use os passos da coluna *X_passo3_quadrado* em vez da coluna *X_passo3*.
- Representamos ambas as colunas, *X_passo3_quadrado* e *Y_passo3*, num gráfico de dispersão.

Tarefa 5

Passo 4

Por último, vamos criar dados com relações diferentes de uma relação linear ou polinómia:

- Podemos usar uma nova folha ou documento, se pretendermos.
- Criamos uma coluna *X_passo4*, com valores na classe *[0, 10, 20, …, 100]*.
- Criamos outra coluna, *Y_passo4*, e calculamos os seus valores em função da fórmula:
  *Y_passo4 = 3 * log(X_passo4 + 2)*
- Representamos ambas as colunas num gráfico de dispersão.

ESPECIALIZAÇÃO EM MACHINE LEARNING

Tarefa 6

Vamos conhecer-nos! E ver como podemos entrar em contacto através da plataforma.

Para esta tarefa, devemos enviar uma mensagem, através da plataforma com a informação seguinte:

- Procuramos informação sobre a frase seguinte e analisamos. Explicamos por que é interessante recordá-la, sempre, ao analisarmos dados. A frase é:
  "A correlação não implica causalidade".
- Procuramos alguns artigos, blogs ou semelhantes, onde exponham exemplos de análise de dados estatísticos, onde a correlação não implicou causalidade e enviamos.
- Procuramos algum meme, piada ou citação engraçados com essa frase ou semelhantes. **Há muitos :D!**
- Procuramos a famosa banda desenhada de XKCD sobre essa frase, analisamo-la e enviá-la-emos juntamente com a explicação da piada.

Tokio.

Aprendemos que...

- Até agora, não realizámos machine learning a sério. Limitámo-nos a criar modelos estatísticos, analisá-los e avaliá-los.
- Vimos, como machine learning não é mais do que realizar modelos estatísticos. Durante o curso, veremos que a diferença está na implementação e em usar algoritmos mais avançados.
- Vimos a necessidade de analisar dados, antes de pretender treinar um modelo.
- Descobrimos o conceito de resíduos.
- Vimos como, muitas vezes, devemos transformar os dados, antes de treinar um modelo sobre eles.
- Aprendemos a criar os nossos próprios datasets sintéticos, para testar modelos e as suas implementações.