UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA
CURSO DE MESTRADO/DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO

LISTA DE EXERCÍCIOS 1 – REGRESSÃO

Os exercícios abaixo referem-se a base de dados "Risco de ataque cardíaco"

1) Obtenha o modelo utilizando os 10 primeiros exemplos da base de dados. Calcule e apresente o erro quadrático médio aplicando o modelo de regressão nos mesmos 10 primeiros exemplos da base de dados. Depois calcule e apresente o erro quadrático médio do modelo de regressão obtido nos demais exemplos. Argumente se o modelo tem ou não uma boa capacidade de predição em novos exemplos.

2) Agora obtenha o modelo utilizando os 5 primeiros exemplos da base de dados e também os 5 últimos. Calcule e apresente o erro quadrático médio aplicando o modelo de regressão nos 10 exemplos utilizados para obter o modelo de regressão. Depois calcule e apresente o erro quadrático médio do modelo de regressão obtido nos demais exemplos. Argumente se o modelo tem ou não uma boa capacidade de predição em novos exemplos. Compare com os resultados do exercício anterior e argumente as possíveis diferenças de resultados.

Os exercícios 3, 4 e 5 referem-se a base de dados "Preços de apartamentos"

3) Qual o coeficiente de correlação entre cada uma das variáveis com o preço de apartamento? Qual a variável mais importante para explicar o preço de apartamento? Justifique sua resposta.
4) O banco de dados contém informações de 40 apartamentos vendidos no mês passado. Cada linha do banco de dados é um apartamento. Ajuste o seguinte modelo de regressão múltipla para os dados:

   $Y=b0+b_1x_1+b_2x_2+b_3x_3+erro$, em que:

   Y=preço do apartamento;
   X1=tamanho do apartamento, em metros quadrados;
   X2=idade do prédio, em anos;
   X3=andar em que o apartamento está.

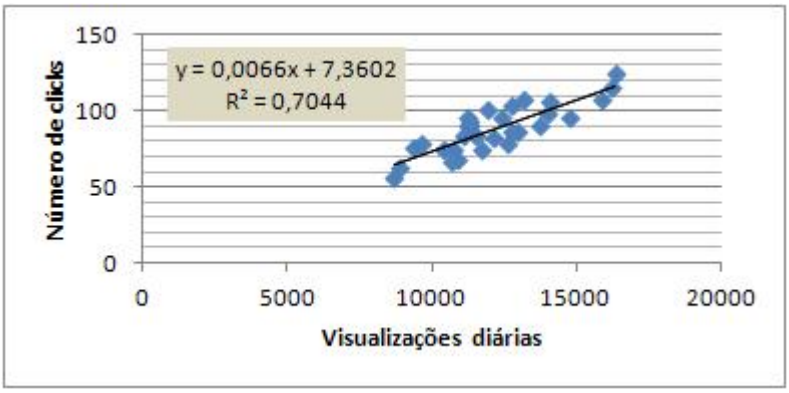
   Obs.: Note que não usaremos todas as variáveis independentes.

   Qual é o preço previsto de um imóvel com 80m2, 10 anos e que está no 9º andar?

5) Ajuste o modelo de regressão múltipla fazendo uso de todas as variáveis. Qual deve ser o preço de um imóvel com 100m2, 3 anos, andar de número 5, 3 quartos e 2 vagas de garagem?

6) Explique com suas palavras a importância do uso da regressão no exemplo deste artigo: https://www.linkedin.com/pulse/using-regression-predict-baseball-salaries-nate-reed

7) Um portal da internet cobra anúncios na página principal de acordo com o número de visualizações diários da página. Um anunciante diz que o mais importante para ele é o número de "clicks" diários no seu anúncio. O portal preparou um estudo com dados dos últimos 30 dias. Foi observado o número de visitas únicas diárias do portal e o número de clicks diários do anúncio. O resultado do ajuste da regressão de Y(número de clicks) por X(número de visualizações) é mostrado no gráfico de dispersão a seguir.



Assinale V ou F justificando sua resposta para as falsas:

(a) Apenas 0,66% das visualizações resultam em clicks. Portanto, a regressão não está boa.
(b) O número de visualizações é um bom preditor do número de clicks porque o R2 é alto e a reta parece bem ajustada.
(c) 0,7044% das visualizações são convertidas em clicks.
(d) As duas variáveis não estão linearmente correlacionadas porque o R2 é menor que 1.
(e) Se num determinado dia o site tiver 10000 visualizações, o número estimado de clicks é 66.

8) Ilustre e explique uma aplicação em que a regressão logística se apresente como mais adequada do que uma regressão linear.

Opcionais:

- Brincando com os coeficientes:

  http://students.brown.edu/seeing-theory/regression/index.html#first

- Notebook de regressão linear para predição de preços de casas na cidade de Boston.

http://facweb.cs.depaul.edu/mobasher/classes/csc478/Notes/IPython%20Notebook%20-%20Regression.html

- Notebook de regressão logística para predição de relacionamento extra-matrimoniais de mulheres:

http://nbviewer.jupyter.org/gist/justmarkham/6d5c061ca5aee67c4316471f8c2ae976