

INTRODUÇÃO AO APRENDIZADO DE MÁQUINA (EST171)

1º SEM./2021

PROF.: THIAGO REZENDE

DEPTO. ESTATÍSTICA – UFMG

DATA DE ENTREGA: 24/06/2021

INSTRUÇÕES:

- 1) O trabalho pode ser em grupo de, no máximo, 3 alunos.
 - 2) Trabalhos em atraso não serão aceitos.
 - 3) É necessária a resolução de cada questão. Quando for pedido o *software python*, coloque um *output* dele (pode ser um *printscreens* da tela) e a análise dos resultados na resolução.
 - 4) A resolução do trabalho e os códigos em python devem ser enviados em formato eletrônico para o e-mail: disciplinas.lst@gmail.com até o dia 24/06/2021. Coloque no assunto do e-mail “TPI – EST171”. É aceito um arquivo do jupyter notebook (formato .IPYNB) que funciona como arquivos *Rmarkdown*, mas é necessário incluir nele os comentários pertinentes.
 - 5) 5 pontos extras para os 3 melhores trabalhos, se houverem.
-

Trabalho Prático I

Instruções Gerais: Quatro bases de dados foram selecionadas para o ajuste de modelos de ML. Procure identificar a distribuição da variável resposta e ajustar o(s) modelo(s) pertinentes, realizando a seleção das variáveis. Inicie com uma análise exploratória e faça considerações sobre os dados. Faça *data wrangling* e a seleção de variáveis. Descreva Separe adequadamente 80% dos dados para treinamento e 20% para validação para os problemas 1 e 4. Para os problemas 2 e 3, use a validação cruzada *K-fold* com K=10 para validar o seu modelo. Faça análise das métricas de desempenho do modelo.

Faça análise das métricas de desempenho do modelo.

Problema 1: (“prostate data”) Deseja-se prever o logaritmo do PSA (Prostate Specific Antigen) com as demais variáveis, usando os métodos de ML introduzidos no curso até agora no python.

Descrição dos dados: These data come from a study that examined the correlation between the level of prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. It is data frame with 97 rows and 9 columns. The data frame has the following components: lcvol log(cancer volume), lweight log(prostate weight), age, lbph log(benign prostatic hyperplasia amount), svi seminal vesicle invasion, lcp log(capsular penetration), gleason

Gleason score, pgg45 percentage Gleason scores 4 or 5 e lpsa log(prostate specific antigen).

Problema 2: (“card data”, Japanese Credit Screening Database) Deseja-se prever a variável binária de saída (Y), ou seja, classificar as instâncias como positiva ou negativa com respeito à concessão de crédito com as demais variáveis, usando os métodos de ML introduzidos no curso até agora no python.

Descrição dos dados: It Includes domain theory: Positive instances are people who were granted credit. The theory was generated by talking to Japanese domain experts. **Credit Card Application Approval Database:** Good mix of attributes -- continuous, nominal with small numbers of values, and nominal with larger numbers of values. 690 instances, 51 input variable, and one output variable.

Problema 3: O conjunto de dados **College**, que pode ser encontrado no arquivo **College.csv**. Ele contém uma série de variáveis para 777 diferentes universidades e faculdades nos EUA. As variáveis são:

- **Private:** Public/private indicator
- **Apps:** Number of applications received
- **Accept:** Number of applicants accepted
- **Enroll:** Number of new students enrolled
- **Top10perc:** New students from top 10% of high school class
- **Top25perc:** New students from top 25% of high school class
- **F.Undergrad:** Number of full-time undergraduates
- **P.Undergrad:** Number of part-time undergraduates
- **Outstate:** Out-of-state tuition
- **Room.Board:** Room and board costs
- **Books:** Estimated book costs
- **Personal:** Estimated personal spending
- **PhD :** Percent of faculty with Ph.D.'s
- **Terminal:** Percent of faculty with terminal degree
- **S.F.Ratio:** Student/faculty ratio
- **perc.alumni:** Percent of alumni who donate
- **Expend:** Instructional expenditure per student
- **Grad.Rate:** Graduation rate

Crie uma nova variável qualitativa binária, chamada **Elite** através da variável **Top10perc**. Divida as universidades em dois grupos com base na proporção de alunos vindos dos 10% melhores classes do ensino médio.

Deseja-se prever a variável binária de saída (Y), ou seja, classificar as instâncias/observações como sim ou não na elite das universidades com o auxílio das demais variáveis preditoras/features, usando os métodos de ML introduzidos no curso até agora no python.

Problema 4: Este problema envolve o conjunto de dados de habitação de **Boston** no EUA. Deseja-se prever o preço com as demais features e entender a

importância de cada *feature* na predição do preço, usando os métodos de ML introduzidos no curso até agora no python.

Descrição dos dados:

Boston house prices dataset.

:Number of Instances: 506.

:Number of Attributes: 13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

:Attribute Information (in order):

- **CRIM:** per capita crime rate by town
- **ZN:** proportion of residential land zoned for lots over 25,000 sq.ft.
- **INDUS:** proportion of non-retail business acres per town
- **CHAS:** Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- **NOX:** nitric oxides concentration (parts per 10 million)
- **RM:** average number of rooms per dwelling
- **AGE:** proportion of owner-occupied units built prior to 1940
- **DIS:** weighted distances to five Boston employment centres
- **RAD:** index of accessibility to radial highways
- **TAX:** full-value property-tax rate per \$10,000
- **PTRATIO:** pupil-teacher ratio by town
- **B:** $1000(B_k - 0.63)^2$ where B_k is the proportion of blacks by town
- **LSTAT:** % lower status of the population
- **MEDV:** Median value of owner-occupied homes in \$1000's

:Missing Attribute Values: None

:Creator: Harrison, D. and Rubinfeld, D.L.

