PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Supervised Machine Learning: Modelos Logísticos Binários e Multinomiais I
Data: 24/08/2021

Rogério Assiz
O script do professor está com algumas diferenças nos números das linhas dos códigos.... a partir da linha 154.... me parece que tem 2 instruções a mais no nosso script
Resposta: O último script disponibilizado aos alunos, na tarde do dia 24/08, é o mesmo utilizado durante a aula.

Marleide Ferreira Alves
o que a função @ faz?
Resposta: O termo @ extrai ou substitui o conteúdo de um slot em um objeto com uma estrutura de classe formal. No caso da aula, extraímos o conteúdo de y.values contido no objeto predicoes.

Murilo Marin Pechoto
A matriz de confusão não deveria ser feita com dados de teste? ou seja, dados não utilizados para treinamento do modelo
Resposta: Pode ser. Se o objetivo é a validação de modelos, pode-se utilizar a matriz de confusão para grupos de treino e validação.

Guilherme Piva Magalhaes Da Rocha
Temos que fazer o teste de Teste de Box-Tidwell para ver se tem relaçao linear entre var X continua e o logito da Var Y?
Resposta: Boa pergunta. Sim, pode-se testar a não linearidade no modelo de regressão logística pelo teste de Box-Tidwell. Sugiro consultar a função box.tidwell(y, ...) no R para computar transformações não-lineares nas variáveis preditoras.

Raphael Fidelis Valadares
Mas o senhor disse que não é necessário que o banco de dados esteja balanceado o que, para mim, é um pouco surpreendente. Se eu sei de antemão que a probabilidade de ocorrência de meu evento é, por...
Resposta: Boa pergunta. De fato o dataset não precisa estar balanceado entre “1” e “0” na variável dependente para modelos logísticos binários. Discutiremos um pouco mais sobre este fato na próxima aula.

Guilherme Piva Magalhaes Da Rocha
Posso usar o Binary Cross Entropy ao inves do Likelihood
Resposta: Sim, para um classificador binário, como é o caso dos modelos logísticos binários, pode-se utilizar a função de custo Binary Cross Entropy, também conhecida por Log Loss, cuja expressão é muito parecida com a expressão de Log Likelihood decorrente da função densidade de probabilidade da distribuição Bernoulli.

Eduardo Garbin
Posse ter um BIC menor e um AIC maior? Qual levo em conta na decisão?
Resposta: Só serão utilizados para comparação com BIC e AIC de outros modelos. Não tem sentido compará-los para um mesmo modelo.

Cainã Max Couto Da Silva
Sobre dados desbalanceados, há problema de desbalanceamento nas variáveis preditoras (X), quer seja nas regressões lineares ou logísticas?
Resposta: Não há problema algum. Discutiremos um pouco mais sobre este fato na próxima aula.

Leonardo Alves Peixoto
Professor, teria como testar se a diferença entre duas AUROC sáo significativas ?
Resposta: Boa pergunta. Esta comparação pode ser realizada pelo teste de DeLong, por meio da função roc.test do pacote pROC. Mostraremos este procedimento na próxima aula.

Alexandro Correa Gonçalves Afonso
variável X causa na probabilidade de ocorrência do evento estando tudo o mais constante (ceteris paribus)? ex: no caso do atraso, a variação da qtd de semáforos é mais sensível que a outra.
Resposta: Sim, está correta a análise. Só peço que não condicione ao termo causalidade.

Thiago Ricardo
Balanceamento de classes não é importante em nenhum modelo de classificação? Nem mesmo em um Naive Bayes?
Resposta: Em algoritmos como Naive Bayes, o preditor negligencia classes em detrimento de outras em bases fortemente desbalanceadas. Como discutido em aula. técnicas como regressão logística não sofrem com tal problema.

Thiago Ricardo
Então balanceamento de classes é irrelevante pra modelos logísticos?
Resposta: Sim. O balanceamento é irrelevante para modelos logísticos. Discutiremos um pouco mais sobre este fato na próxima aula.