PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Supervised Machine Learning: Modelos Logísticos Binários e Multinomiais II
Data: 31/08/2021

Guilherme de Castro Ribeiro Luz
É possível trabalhar com F1 score como métrica de avaliação de performance do modelo multinomial?
Resposta: Sim, este procedimento tem sido feito para escoragem de observações após a estimação do modelo.

Gabriel Campos Vieira
Teria algum teste similar ao lrtest para comparação de modelos glm?
Resposta: O Lagrange multiplier (ou score test) também é utilizado para um único modelo. Para se compararem estimações provenientes de diferentes modelos, o mais utilizado mesmo é o lrtest.

Gabriel Campos Vieira
Corrigindo a pergunta: Teria algum teste similar ao lrtest para comparação de modelos lineares ou com transf box-cox??
Resposta: Entendi a pergunta. Se os modelos forem estimados por máxima verossimilhança, o mais correto é que suas estimações sejam comparadas a parti do lrtest.

Eduardo Luís Hammes
Rodei o código mas removendo a Dummy mais frequente e a Área ROC diminuiu. Existe uma maneira de remover a dummy que maximize essa área?
Resposta: Se o beta da referida dummy for estatisticamente significante, a área abaixo da ROC nunca será aumentada ao se eliminar aquele beta do modelo. A única maneira de se aumentar a área abaixo da ROC é a inclusão de parâmetros significantes, seja no componente de efeitos fixos, seja no componente de efeitos aleatórios.

Cristiano Domingos Gonçalves Gomes
Boa noite Professor! Dentre os métodos de diagnóstico, em algum momento vai ser ensinado métodos de validação cruzada como k-fold, por exemplo.
Resposta: K-fold é bastante utilizado como um procedimento de reamostragem para se avaliarem modelos em amostras limitadas. Sugiro o estudo de “An Introduction to Statistical Learning: with Applications in R”.

Myke Morais De Oliveira

dividir dataset em teste e treino não deve fazer sentido em modelos GLM né?

Resposta: É uma boa pergunta. Pode ser feito (e é feito no mercado), mas como a estimação dos parâmetros é determinística e definida por máxima verossimilhança, acreditamos que não faz muito sentido que os parâmetros sejam estimados e o algoritmo seja treinado em um subset menor para fins preditivos (se este for o caso, a razão é para verificação de aderência e consistência dos dados, mas não para fins preditivos para outras observações não presentes inicialmente no banco de dados). Em técnicas com estimação estocástica, como alguns ensemble e redes neurais, faz mais sentido.

Rick Tavares Barbosa

Para quais situações eu preciso separar uma amostra de treino e teste?

Resposta: É uma boa pergunta. Pode ser feito (e é feito no mercado), mas como a estimação dos parâmetros é determinística e definida por máxima verossimilhança em modelos GLM, acreditamos que não faz muito sentido que os parâmetros sejam estimados e o algoritmo seja treinado em um subset menor para fins preditivos (se este for o caso, a razão é para verificação de aderência e consistência dos dados, mas não para fins preditivos para outras observações não presentes inicialmente no banco de dados). Em técnicas com estimação estocástica, como alguns ensemble e redes neurais, faz mais sentido.

Jean Carlos Zambrano Contreras

Na seleção das variáveis dos modelos como posso saber qual método usar forward ou backward, condicional...?

Resposta: A estimação dos parâmetros converge para os mesmos resultados. O procedimento Stepwise já faz esta análise.

Guilherme Piva Magalhaes Da Rocha

Qual a diferença entre likelihood e Binary cross entropy na regressao logistica binaria?

Resposta: A lógica é a mesma e a expressão de ambas é similar também.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

Leandro Modesto Prates Beltrão

É possível o beta de uma mesma variável ser estatisticamente significante para um logito, mas não significante para outros?

Resposta: Ótima pergunta. Sim, com certeza o beta de determinada variável X pode ser estatisticamente significante em um logito e não ser em outro logito para uma regressão logística multinomial.

Damião Flávio dos Santos

Quero estimar um modelo de regressão linear múltipla (y contínuo) com variáveis x multinomial, mas há a presença de NA, seria uma boa tentar estimar os NAs usando o modelo logístico multinomial? Obg

Resposta: É uma boa pergunta, com a qual concordo. Já fizemos isso algumas vezes em modelagens com esta característica no passado.

Paulo Renato Leite

Professor, tanto o lrtest quanto o roc.test comparam somente 2 modelos?

Resposta: Sim, a comparação se dá entre dois modelos.

Marcelo Matsumoto

Professor, quando a distribuição dos valores nas variáveis preditoras (x) não tem um padrão de distribuição gaussiana, a aplicação da Random Forest não seria melhor do que uma GLM?

Resposta: Embora se diga isso por aí, não há pressuposto algum de que as variáveis preditoras precisam ter distribuição gaussiana. Se este fosse o caso, nunca poderíamos utilizar dummies como variáveis X. Na realidade, podemos inclusive comparar os fitted values de modelos GLM com modelos do tipo random forest, por exemplo.

André Sigora

Professor, qual sua opinião sobre as f-measures (F1-score, F0.5, F2) para medir performance dos modelos?

Resposta: *F-measure* é uma combinação das métricas de *precision* (quanto maior a *precision*, menor o número de falsos positivos - FP) e sensitividade (taxa de verdadeiros positivos - TP) em uma matriz de confusão para dado *cutoff*. Representa outro indicador de performance de modelo para dado cutoff. Um alto valor de *F-measure* garante que a *precision* e a sensitividade sejam razoavelmente altas. A sua expressão é:

$$F = \frac{2 \times TP}{2 \times TP + FP + FN}$$