PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

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

Gabriel Rodrigues Coutinho Pereira
Não existe curva ROC para modelos de regressão logística multinomial?
Gabriel, a curva *ROC (Receiver Operating Characteristc)* é uma ferramenta que permite avaliar o desempenho de um modelo de regressão binária (quando a variável resposta é do tipo evento/não evento). Ou seja, a curva ROC é um gráfico de Sensibilidade (ou taxa de verdadeiros positivos) versus taxa de falsos positivos, ou seja, representa a Sensibilidade (ordenadas) *versus* 1 – Especificidade (abcissas) resultantes da variação de um valor de corte. Portanto, não há curva ROC para modelos multinomiais. Esta questão se dá em virtude do modo de sua construção, em que são discriminados acertos para os valores considerados positivos e para os valores considerados negativos. Esta classificação não é possível para modelos multinomiais.

WAGNER ELIAKIM LUZ LIMA
Chegar atrasado na segunda aula, implica em chegar atrasado na primeira aula. Seria um erro de modelagem no dado, considerar esses dois casos ocorrência de evento (1 e 1)?
O seu raciocínio está correto, Wágner. É um outro jeito de se pensar essa análise. A decisão sobre o modelo de análise depende do objetivo de pesquisa. É preciso pensar no porquê de se discriminar os grupos de pessoas que chegam atrasadas. É necessário incluir mais uma classe? É necessário se distinguir entre "graus de atraso"? É fundamentalmente uma decisão de pesquisa.

Ivan Cesar Desuo
Como fica a questao do stepwise no modelo multinomial? Posso aplicar como nos outros modelos?
Funciona de modo semelhante nos outros modelos, Ivan. Baseado em parâmetros de escolha definidos pelo algoritmo se retiram as variáveis não significativas para a análise.

Bernardo Silva
Prof. no caso da multinomial também aplica-se o ajuste do alfas quando há conhecimento do comportamento dos valores da Y no universo de evento estudado?
Exatamente, Bernardo. Com o prévio conhecimento do comportamento dos valores para a população pode-se proceder ao ajuste do alfa.

Raphael Fidelis Valadares
Professor, você sabe dizer se existe alguma biblioteca de Python que faça regressões e stepwise? O statsmodels faz regressões de todos os tipos, mas não faz stepwise...
Raphael, desconheço tal teste. Peço a gentileza de que nos lembre no e-mail da monitoria para que possamos pesquisar sobre essa questão e tentar te ajudar.

Daniela Maciel Pinto

Quando rodo o código: dados_plotagem - cbind.data.frame(cutoffs, especificidade, sensitividade) recebo o erro: Error in data.frame(..., check.names = FALSE) : objeto 'cutoffs' não encontrado

Daniela, provavelmente algum comando acabou não sendo executado no R, por diversos possíveis motivos. Nos encaminhe sua dúvida no e-mail da monitoria, com um print do erro, se possível. Iremos prontamente lhe auxiliar.

Marcelo Patto

rodei dummie (linha#382, depois stepwise (linha#411), mas o summary nao mudou, preço03 ainda nao passa! o que errei?

Marcelo, nos encaminhe sua dúvida no e-mail da monitoria, com um print do erro, se possível. Iremos prontamente lhe auxiliar.

Carlos Henrique de Oliveira

Por que em alguns casos utilizamos $ e outros @ para acessar estes compartimentos? Existe uma lógica específica, por exemplo objetivos do tipo A utilizamos $ e do tipo B utilizamos @?

Carlos Henrique, são os operadores característicos do R. O $ serve para especificar uma variável dentro de uma base e o @ elenca alguns outros objetos dentro das funções.

Larissa Ribeiro Rios

Pode explicar novamente sobre o Log Lik, por favor?

O Log Likelihood é uma medida de adequação de um modelo estatístico. Quanto maior o valor, melhor é o modelo. Larissa, como você quer maximizar o Log Likelihood , o maior valor é melhor. Por exemplo, um valor de log-lik de -2 é melhor que -10. Larissa, caso ainda não tenha ficado claro, peço que nos acione no e-mail da monitoria, iremos te auxiliar do melhor modo possível.

Vitor Bruno da Silveira Guimarães

O LogLik, quanto menor é melhor?

Vitor, como você quer maximizar o log-lik, o maior valor é melhor. Por exemplo, um valor de Log Likelihood de -2 é melhor que -10.

Marcelo Patto

rodei stepwise apos dummização e preço3 nao passou

Marcelo, nos encaminhe sua dúvida no e-mail da monitoria, com um print do erro, se possível. Iremos prontamente lhe auxiliar.

Ivan Cesar Desuo

Possuo N variáveis X Quali binárias (SIM/NÃO) no meu banco de dados. Por serem binária e cada observacao poder assumir S/N para todas as N variáveis. Devo ou não dumizar pelo n columns?

Ivan, se bem entendi seu objetivo, a intenção é especificar a presença ou ausência de determinado sintoma em determinada pessoa. Nesse caso teríamos várias colunas com os diversos sintomas especificando a presença ou ausência daquele sintoma em específico para cada observação. Caso tenha entendido corretamente, sim, esse procedimento de dummyzação precisa ser realizado. Caso ainda tenha dúvidas, peço que acione a monitoria pelo e-mail que será prontamente atendido.

Vitor Bruno da Silveira Guimarães

Com a dummyzação é possível utilizar em outros modelos?

Sim, Vitor. A dummyzação permite a utilização desta variável em modelos diversos.

Cleverson de Souza

Como eu uso o predict para gerar a probabilidade de um novo cliente?

Cleverson, de posse dos dados desse cliente você coloca estes valores nos parâmetros correspondentes, estimando assim a probabilidade para aquelas características específicas. Lembrando sempre da questão da não extrapolação quando for realizar previsões.

Alessandro Nogueira Corrêa

Por que se trabalha com o logaritmo da chande(odds) no modelo de regressão logística em vez de se utilizar apenas a probabilidade?

Alessandro, as duas possibilidades são possíveis a depender do seu objetivo de pesquisa. Posso estar interessado no quanto uma determinada dose de remédio aumenta a probabilidade de não ter determinada doença para determinada pessoa, por exemplo. Ou posso querer investigar em quanto a chance daquela pessoa ter determinada doença diminui caso o remédio seja administrado. São duas formas de investigação plenamente possíveis. Do ponto de vista matemático são questões análogas.

Cleverson de Souza

Onde eu aplico esse modelo obtido?

Cleverson, é um modelo adequado para estimação de probabilidades. De posse destes parâmetros é possível calcular as probabilidades para outros indivíduos da população. Estamos criando um algoritmo para prever comportamentos de indivíduos semelhantes que não constam dessa amostra inicial, bem como verificando o impacto das variáveis preditoras no fenômeno que estamos estudando.

Marcelo Patto

Prof. depois do intervalo poderia explicar quantidade de dummies novamente por variavel plz?

Marcelo, se entendi bem sua pergunta, a quantidade de dummies é n-1. Onde “n” é a quantidade de categorias de cada variável. Caso tenha uma variável, por exemplo, com as seguintes opções de diferencial semântico: “discordo”, “neutro”, “concordo”; teremos aí três categorias, sendo, portanto, um indicativo de duas dummies.

Anderson de Oliveira Pinto
Como eu digo pro modelo que ele deve me devolver a probabilidade de falha? e se eu quisesse a probabilidade de "não falha"? como faria?
Anderson, essa questão é feita por meio do argumento “ref” da função “relevel”. Esta especificação é melhor discutida pelo professor Fávero na aula “Supervised Machine Learning: Modelos Logísticos Binários e Multinomiais II” a partir das 3 horas e 8 minutos da aula.

Leonardo Moreira Paes De Camargo
Professor, eu entendi a visualização, mas eu não acredito que se eu apresentasse para a minha Diretoria eles entenderiam. Qual seria a sua dica para que eu pudesse fazê-los entnder de 1 forma simples?
Leonardo, é sempre um desafio. Creio que a melhor forma de apresentação seja uma apresentação em função dos coeficientes da regressão. Dado um aumento em determinada variável, qual o incremento na variável dependente. Aumento a quantidade investida em promoção, quanto eu consigo de retorno? Aumentando o investimento em prospecção de clientes, em quanto eu consigo aumentar a probabilidade de compra? São construções interessantes para se apresentar aos gestores, mas realmente é uma tarefa que tem alguns desafios.

Samya de Lara Lins de Araujo Pinheiro
Como interpretar os betas no sentido de captar mudança na probabilidade (ou chance/odds ratio) para uma mudança de uma unidade na variável X (ex aumento em 1unidade para a variável distância) ?
Samya, o raciocínio é o seguinte: em média, em quanto se altera a probabilidade de se chegar atrasado à escola ao se adotar um percurso 1 quilômetro mais longo, mantidas as demais condições constantes? Essa é a pergunta que estamos tentando responder na regressão logística.

Isac Terceiro
Professor, poderia falar sobre a diferença entre logit e probit? e se vai abordar probit nas aulas.
Isac, a logística possui caudas um pouco mais achatadas, isto é, a curva probit se aproxima dos eixos mais rapidamente que a curva logit. Os modelos logit e probit, no entanto, são apenas *modelos*. Ambos os modelos permitirão detectar a existência de um efeito de no resultado; exceto em alguns casos muito especiais, nenhum deles será "realmente verdadeiro", e sua interpretação deve ser feita com cautela e levando-se em consideração as características de cada problema de pesquisa e banco de dados. Caso haja maior preocupação com a parte final da curva, em algum momento a seleção do logit ou probit será importante. Não existe uma regra exata para selecionar probit ou logit. Você pode selecionar o modelo observando os indicadores de ajuste de cada um dos modelos estimados.

Leonardo Moreira Paes De Camargo
Era do exemplo que deu 16 dummies
Leonardo, se entendi bem sua pergunta, eram 4 variáveis com 5 categorias, nesse caso teríamos n-1 dummies para cada uma das variáveis, perfazendo, portanto, um total de 16 dummies (4 * (n-1)) = (4 * (5-1)) = 4*4 = 16.

Leonardo Moreira Paes De Camargo
sao 5 ou 4 categorias?
Leonardo, se entendi bem sua pergunta, eram 4 variáveis com 5 categorias, nesse caso teríamos n-1 dummies para cada uma das variáveis, perfazendo, portanto, um total de 16 dummies (4 * (n-1)) = (4 * (5-1)) = 4*4 = 16.

Alexandre Gonçalves da Rocha
Desculpem minha ignorância, talvez tenha perdido algo, mas lá vai: se não existe R quadrado para modelos com variáveis qualitativas, por que o R apresenta o valor?
Alexandre, esse indicador existe e é tratado na literatura. No entanto, do mesmo modo, há algumas ressalvas em relação à sua utilização para modelos logísticos. Essa questão foi melhor tratada pelo professor Fávero na aula "Supervised Machine Learning: Modelos Logísticos Binários e Multinomiais I".

Gabriel Trazzi
Professor, utilizar uma categorização de um y numérico (mais de 3 categorias, por exemplo) cujo modelo não ficou legal, seria uma opção tecnicamente adequada, aplicando o modelo na forma multinomial?
Gabriel, se entendi sua pergunta, você está propondo categorizar números em determinadas faixas. Por exemplo, caso o valor numérico de idade não seja significativo em determinada análise, transformar para uma variável que apresente as faixas de idade. Nesse caso, sim. É plenamente possível.

Luca Samuel Ghinaim Moreto
Então quando as variáveis X forem binárias eu posso fazer uso da dummização automática ?
Luca, sim. O R nos oferece essa possibilidade para variáveis qualitativas.

Vanessa Hoffmann de Quadros
Professor, poderia comentar um pouco sobre como interpretar os betas no modelo logit?
Vanessa, o coeficiente estimado para uma variável preditora representa a mudança na função de ligação para cada mudança de unidade na preditora, enquanto as outras preditoras no modelo são consideradas constantes.