Autor (a): Karla Guimarães S. Pereira

Proposta:

¨No desenvolvimento de modelos de predição, qual a diferença entre as técnicas de regressão linear e regressão logística? Quais são os indicadores para avaliar a performance de aderência do modelo?"

Para o desenvolvimento destas respostas requeremos revisar de modo superficial alguns critérios:

Modelo de regressão linear:

A regressão linear possibilita o estudo da relação entre uma ou mais variáveis explicativas. Tem o objetivo de avaliar o comportamento da variável Y em função do comportamento de uma ou mais variáveis X, sem que necessariamente haja uma relação de causa e efeito (Fávero et al, 2009). A função pode ser definida por:

$$Y_i = a + b_1X_1 + b_2X_2 + \cdots + b_kX + u$$

Em que: Y é o fenômeno a ser estudado (variável dependente quantitativa), a o intercepto, b coeficientes de cada variável (coeficiente angular), X as variáveis explicativas e u o termo de erro (diferença entre o valor real de Y e o valor previsto de Y, para cada observação).

O Método dos Mínimos Quadrados (MMQ) é uma técnica clássica para se estimar modelos de regressão é um estimador que minimiza a soma dos quadrados dos resíduos da regressão.

Segundo Wooldridge (2012), a existência do termo de erro está relacionada a falhas na especificação do modelo ou ocorrência de erros no levantamento de dados. Sendo assim o critério do MMQ são: a soma dos erros deve ser igual a zero e a somatória dos erros ao quadrado seja a menos possível.

Indicadores de performance:

O R^2 é um indicador de percentual de variância da variável Y devido ao comportamento das variáveis explicativas X. O coeficiente varia de 0 a 1, quanto mais próximo de 1 maior é o poder preditivo do modelo de regressão.

Com o intuito de avaliar a significância estatística do modelo de regressão após cálculo do R^2 realiza-se os testes:

> *Teste F:* analisa se pelo menos um dos b são estatisticamente significante para explicação do comportamento de Y.
>
> *Hipóteses:* $H_0: b_1 = b_2 = b_3 = \cdots = b_k = 0$ $\quad H_1: pelo\ menos\ um\ b\ \neq 0$
>
> Se F < 0,05, existe pelo menos um $b \neq 0$
>
> *Teste t*: analisa individualmente se cada um dos parâmetros é diferente de zero:

Hipóteses: $H_0: b = 0$ $H_1: b \neq 0$

Com o valor de t, deve-se utilizar a tabela de distribuição para obtenção dos valores críticos a um dado nível de significância e verifica-se se rejeita-se ou não a hipótese nula.

Modelo de regressão logística:

A Regressão Logística tem o intuito de estimar a probabilidade associada a ocorrência de determinado evento Y. É aplicada quando a variável dependente é dicotômica ou binária.

$$Z_i = a + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_{ki} X$$

Em que Z é o logito, a uma constante, b parâmetros estimados de cada variável explicativa e X variáveis explicativas (métricas ou dummies)

A ocorrência de um evento é denominado Chance(odds): $Chance = \frac{p}{1-p} \frac{(evento)}{(não\ evento)}$

A sigmoide descreve a relação entre a probabilidade associada à ocorrência de determinado evento e um conjunto de variáveis preditoras. A função logística assume valores entre 0 e 1 para qualquer Z entre $-\infty\ e\ +\infty$.

$$p_i = \frac{1}{1 + e^{-Z}}$$

Indicadores de performance:

A estimação do modelo de regressão logística é realizada pelo Método da Máxima Verossimilhança e assim como regressão linear, devemos calcular R^2 e as significâncias estatísticas de Z, seguido de matriz de confusão.

O cutoff é um ponto de corte a ser definido na amostra a ser classificada como evento. Vale lembrar que a definição de cutoff é válida apenas para regressão logística binária.

Com os valores da matriz de confusão devemos calcular a eficiência global do modelo, sensitividade e a especificidade e em seguida plotar a curva ROC que apresenta a variação da sensitividade em função de (1 – especificidade).

Qual a diferença entre as técnicas de regressão linear e regressão logística

No critério de classificação a regressão logística fornece resultados estatísticos superiores se comparado a regressão linear. No setor bancário a regressão logística é utilizada em modelos de prevenção de fraude de cartão de crédito (Beraldi F., 2014). No setor tributário empregada para calcular a probabilidade do contribuinte ser inadimplente ou adimplente após o parcelamento de tributos (Dias F., 2003). Na medicina, classificação de indivíduos doentes e sãos.

A técnica supervisionada de regressão logística é comparada aos modelos de árvore de decisão, redes neurais, SVM. etc.

Modelos de regressão linear são aplicados para prever determinado fenômeno, como por exemplo no cálculo de faturamento do segundo semestre de 2022 de um determinado supermercado com base no faturamento do primeiro semestre.