Técnicas de Machine Learning II

Objetivo: Analisar uma variável de interesse, a variável de target, explicando o comportamento da variável Y, em função de variáveis X, definidas genericamente como variáveis x que são as variáveis preditoras, que esperamos que sejam relevantes na explicação do comportamento de Y. As variáveis X também podem ser chamadas de features, de características do problema em questão, ou seja, são variáveis que explicam o comportamento de Y. Estes são os modelos supervisionados, porque propomos as relação a ser testada, a relação a ser modelada, determinando Y em função de X gerando um modelo com funções preditivas.

* A variável Y, ou seja a variável a ser explicada, também chamada de variável dependente, é sempre uma variável do tipo métrica.
* Sobre a variável X podemos ter um conjunto de variáveis X, e elas podem ser tanto métricas quanto categóricas.
* Modelos de regressões simples ou múltiplas
  + Simples: Apenas 1 Variável X
  + Múltiplas: Mais de uma variável X
* Gráfico de scaterPlot:

Gráfico de Pontos

* RegPlot:

ScaterPlot mais rebuscado com mais rebuscado com mais informação usado para regressão linear plotando também o ajuste Linear.

* Mínimos quadrados Ordinários
* Estatística F TESTE F (ANOVA)
  + Avalia a significância geral do modelo de regressão, ou seja, se pelo menos um dos β estimados é diferente de zero.
* SQR = Soma dos quadrados da Regressão
* SQU = Soma dos quadrados dos resíduos
* K = nº de parâmetros do modelo (inclui α)
* N = Tamanho da Amostra
* H0: β\_1=β\_2=⋯=β\_k=0
* H1: Existe pelo menos um β\_j ≠ 0
* Normalmente adota-se o nível de significância de 5% para o teste
* Se o p-valor do teste F <0,005 Rejeitas -se H0
* Ou seja der o p-Valor do teste for menor que 0,05 eu aceito o modelo
* SQT (Soma dos quadrados Totais)
  + SQT = SQR + SQU
    - Soma dos quadrados Totais (SQT): Variação de Y em torno da sua média
    - Soma dos quadrados da regressão (SQR): Variação de Y considerando as variáveis X
    - Soma dos quadrados dos Resíduos(SQU): variação de Y que não é explicada pelo modelo

Esquemático

Descrição gerada automaticamente

Anova = Análise de Variância

* Teste T
  + Avalia a significância individual dos parâmetros estimados

Tela de celular com texto preto sobre fundo branco

Descrição gerada automaticamente com confiança média

* + Adota-se o nível de significância de 5% para o teste
    - Se o p-valor do teste t < 0,05, rejeita-se H0
    - **Mesmo que não tenha significância o α não deve ser removido do modelo**

Fim Bloco 1

Assistir curso álgebra linear mat3457

<https://www.youtube.com/watch?v=-JcQJFNVjaA&list=PLIEzh1OveCVczEZAjhVIVd7Qs-X8ILgnI>

Bloco 2

* Coeficiente de explicação (R²)
  + O R² apresenta o poder explicativo do modelo, ou seja, o percentual da variabilidade de Y que é explicado pela variação das variáveis X
  + O R² varia ente 0 e 1 :Valores mais próximos de 1 indicam maior capacidade preditiva (0 é Ruim 1 é Ótimo)
  + O R² não deve ser analisado no sentido de validar ou não o modelo, pois em muitos campos do conhecimento é comum não obter valores muito elevados
  + O R² ajustado para comparação entre modelos
    - * Ajusta-se a quantidade K de parâmetros (incluindo α) e o tamanho da amostra n
* Fitted Values
  + Intervalos de confiança
    - Para o nível de confiança escolhido, é o intervalo de valores que contém o verdadeiro parâmetro populacional

Texto

Descrição gerada automaticamente

* + - T é o valor critico bicaudal da distribuição t de student para o nível de confiança escolhido na analise, com n-k graus de liberdade
    - Normalmente, observa-se o nível de confiança de 95% (Nível de significância de 5%)
    - Se o intervalo de confiança contem o 0 não são valores estatisticamente relvantes.
* Parâmetros do modelo
  + Interpretação dos parâmetros α e β
    - α é o coeficiente Linear, ou seja, o valor de Y caso todas as X = 0
      * Muitas vezes o α pode ser interpretado como a projeção da reta no eixo Y uma vez que não se encontram-se observações da amostra com todas as variáveis X=0
    - β são os coeficientes angulares, ou seja, a inclinação da reta
      * Na regressão múltipla os β são interpretados na condição ceteris paribus, ou seja o efeito daquela X sobre Y **mantidas todas as demais variáveis constantes.**
      * Destaca-se que a interpretação dos parâmetros do modelo deve ocorrer sem extrapolação dos dados, isto é, vale dentro do limite de variação das variáveis.
    - Devemos respeitar os limites de interpolação do modelo quando realizamos previsões de fora da amostra não fazendo extrapolação dos dados. Devemos respeitar o limite de variação de X ou seja dentro da variação do mínimo e do máximo observado.

Fim bloco 2

Bloco 3

* Coeficiente de explicação (R²)
  + O R² apresenta o poder explicativo do modelo, ou seja, o percentual da variabilidade de Y que é explicado pela variação das variáveis X
  + O R² varia ente 0 e 1 :Valores mais próximos de 1 indicam maior capacidade preditiva (0 é Ruim 1 é Ótimo)
  + O R² não deve ser analisado no sentido de validar ou não o modelo, pois em muitos campos do conhecimento é comum não obter valores muito elevados
  + O R² ajustado para comparação entre modelos
    - * Ajusta-se a quantidade K de parâmetros (incluindo α) e o tamanho da amostra n
    - K = quantidade de parâmetros = variáveis + α
      * Serve para comparação de modelos
* Variáveis explicativas categóricas
  + Quando há variáveis categóricas é necessário transformá-las em dummies
    - Dummy : Variável binaria (1ou0) indicando a presença ou ausência do atributo.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Id | Vriavel A | | Variavel B | | |
| Categ 1 | Categ 2 | Categ 1 | Categ 2 | Categ 3 |
| 1 | 1 | 0 | 0 | 1 | 0 |
| 2 | 0 | 1 | 0 | 0 | 1 |
| 3 | 0 | 1 | 1 | 0 | 0 |
| 4 | 1 | 0 | 0 | 0 | 1 |
| 5 | 0 | 1 | 0 | 1 | 0 |
| 6 | 1 | 0 | 1 | 0 | 0 |

Id | Vriavel A | Variavel B

| Categ 1 | Categ 2 | Categ 1 | Categ 2 | Categ 3

--------------------------------------------------------

1 | 1 | 0 | 0 | 1 | 0

2 | 0 | 1 | 0 | 0 | 1

3 | 0 | 1 | 1 | 0 | 0

4 | 1 | 0 | 0 | 0 | 1

5 | 0 | 1 | 0 | 1 | 0

6 | 1 | 0 | 1 | 0 | 0

Na regressão Utiliza-se o procedimento de N-1 dummies, ou seja uma das categorias de cada variável categórica fica como a referência de sua variável no intercepto

* Procedimento de stepwise
* Procedimento para a remoção de variáveis não significativas

Fim