ASSOCIAÇÃO EDUCACIONAL DE VITÓRIA
FACULDADES INTEGRADAS SÃO PEDRO
CURSO DE GRADUAÇÃO EM
SISTEMAS DE INFORMAÇÃO

RICARDO RODRIGUES DE SOUZA

MODELOS DE ANÁLISE DE DADOS

VITÓRIA
2022

RICARDO RODRIGUES DE SOUZA

MODELOS DE ANÁLISE DE DADOS

Trabalho acadêmico do Curso de Graduação em Sistemas de Informação, apresentado às Faculdades Integradas São Pedro como parte das exigências da disciplina Análise de Dados Aplicada a Computação, sob orientação do professor Howard Cruz Roatti.

VITÓRIA

2022

SUMÁRIO

1 REGRESSÃO LINEAR

Um modelo de Regressão Linear visa estimar o valor de uma variável *y* dependente, baseando-se em outras variáveis *x* independentes. A linearidade do relacionamento entre essas variáveis facilita a interpretação. A equação de um modelo de Regressão Linear é a seguinte:

$$Y = \beta 0 + \beta 1x1 + \ldots + \beta pxp + \epsilon$$

O resultado previsto é a soma de seus aspectos *p*. Os betas (βj) representam os pesos, ou coeficientes aprendidos desses aspectos. O épsilon (ϵ) é o erro que ainda será cometido, como por exemplo a diferença entre o valor previsto e o resultado verdadeiro. Esses erros são previstos para seguirem uma Distribuição Gaussiana, o que significa que erros serão cometidos tanto em direções positivas quanto negativas, e que ocorrerão vários erros pequenos e poucos erros grandes.

A maior vantagem de um modelo de Regressão Linear é a sua linearidade: Ela torna o processo de estimativa mais simples, e podem ser facilmente interpretadas em um nível modular. Esse é um dos motivos pelos quais esse tipo de modelo é tão utilizado em áreas acadêmicas, tais como Medicina, Sociologia, Psicologia, e muitas outras áreas que trabalham com dados quantitativos. Por exemplo, no campo da Medicina, é necessário não apenas prever o resultado clínico de um paciente, como também poder quantificar a influência que um certo remédio terá em seu organismo, levando em conta fatores como sexo, idade, entre outros, que influenciam diretamente nas reações que seu corpo terá.

1.1 EXEMPLOS

É possível utilizar um modelo de Regressão Linear para prever, por exemplo, a quantidade de bicicletas à serem alugadas em uma loja, em um dia específico, baseando-se em dados observados/gerados no passado. Poderia se levar em conta variáveis tais como a Estação do Ano, Clima, se esse dia é um feriado ou não, se é um dia útil ou se é fim de semana, qual a temperatura, velocidade do vento etc.

Tomemos como base os dados abaixo:

	Peso	Erro Estimado	Valor absoluto estatístico
(Interseção)	2399.4	238.3	10.1
estacaoPRIMAVERA	899.3	122.3	7.4
estacaoVERAO	138.2	161.7	0.9
estacaoOUTONO	425.6	110.8	3.8
feriadoFERIADO	-686.1	203.3	3.4
diautilDIAUTIL	124.9	73.3	1.7
climaNUBLADO	-379.4	87.6	4.3
climaCHUVA/NEVE/TEMPESTADE	-1901.5	223.6	8.5
temperatura	110.7	7.0	15.7
velocidadeVENTO	-42.5	6.9	6.2

Baseado nos dados acima, podemos realizar as seguintes interpretações:

Analisando uma propriedade numérica (temperatura), podemos concluir que, o aumento de 1ºC na temperatura faria com que a quantidade prevista de bicicletas alugadas aumentaria em 110.7, considerando que todas as outras propriedades permanecessem com o mesmo valor.

Se analisarmos uma propriedade categórica (climaCHUVA), podemos concluir que a quantidade estimada de bicicletas alugadas é -1902.5 menor quando estiver chovendo, nevando ou com tempestade. Novamente, assumindo que todas as outras propriedades permaneçam com o mesmo valor. Quando o clima estiver nublado, a quantidade total de bicicletas alugadas seria -379.4 menor comparado a quando o clima estivesse aberto, novamente considerando que todas as outras propriedades permaneçam com o mesmo valor.

Devido à natureza dos modelos de Regressão Linear, todas as interpretações são feitas levando em conta que “todas as outras propriedades permaneçam com o mesmo valor”. A parte negativa é que a interpretação ignora a distribuição conjunta das propriedades. Alterar o valor de uma propriedade, mas não alterar de outra pode

fazer com que se produza resultados irreais. Um exemplo prático disso seria aumentar a quantidade de cômodos em uma casa sem aumentar o tamanho da casa em si.

2 REGRESSÃO LOGÍSTICA

Um modelo de Regressão Logística visa determinar a probabilidade de um evento acontecer. Ele leva em conta a relação entre as propriedades, e usa disso para calcular um determinado resultado. É similar à Regressão Linear, porém em vez de um resultado gráfico, a variável prevista é binária: 0 ou 1.

Para fins de classificação, é utilizado probabilidades calculadas apenas entre 0 e 1, portanto a equação da Regressão Logística consiste em encapsular o lado direito da equação de Regressão Linear. Isso força o resultado a assumir valores apenas entre 0 e 1:

$$P(y(i)=1)=11+exp(-(\beta 0+\beta 1x(i)1+\ldots+\beta px(i)p))$$

2.1 EXEMPLOS

Usaremos o modelo de Regressão Logística para prever *câncer cervical* baseado em alguns fatores de risco. A tabela a seguir mostra o peso (que cada propriedade influencia) estimado, a proporção de probabilidades, e o erro padrão das estimativas:

	Peso	Proporção de probabilidades	Erro padrão
Interferência	-2.91	0.05	0.32
Contraceptivos hormonais	-0.12	0.89	0.30
Fumante (S/N)	0.26	1.30	0.37
Número de gestações	0.04	1.04	0.10
Número de DST's diagnosticadas	0.82	2.27	0.33
Aparelhos intrauterinos (S/N)	0.62	1.86	0.40

Interpretação de uma propriedade numérica ("Número de DST's diagnosticadas"):

Um aumento no Número de DST's diagnosticadas aumenta as chances de câncer cervical em um fator de 2,27, considerando que todas as outras variáveis permaneçam com o mesmo valor. Lembrando que essa correlação não implica causa.

Interpretação de uma propriedade categórica ("Contraceptivos hormonais"):

Para mulheres que estiverem utilizando contraceptivos, as chances de **câncer x não-câncer** são 0,89 vezes menores comparado a mulheres que não utilizam de

hormônios. Novamente, considerando que todas as outras variáveis permaneçam com o mesmo valor.

Como no modelo de Regressão Linear, todas as interpretações são feitas com a premissa de que “todas as outras variáveis permaneçam com o mesmo valor”.

3 REFERÊNCIAS

- https://christophm.github.io/interpretable-ml-book/logistic.html