Ciência de Dados e Big Data

**Técnicas Estatísticas de Predição**

***Programa***

**Calendário Conteúdo Referencial Teórico** 05/02/2019 Regressão Linear (Simples e Multipla) 12/02/2019 Regressão Logística (Simples e Multipla) 19/02/2019 Modelo Multinomial Categorical Data Analysis, 2013 26/02/2019 Árvores de decisão e Florestas Aleatórias 12/03/2019 KNN, PCA e K Means Clustering 19/03/2019 Avaliação -An Introduction to Statistical Learning with Applications in R, 2013 An Introduction to Statistical Learning with Applications in R, 2013

***Regressão Logística***

**Livro texto:** *An Introduction to StatisticalLearning*

Cap. 4 – Classification

4.3 Logistic Regression

Usualmente no contexto de *DS* encaramos a regressão logística como um método de classificação.

Exemplos de problemas que desejamos classificar:

• *Filtro de e-mails* (*Spams ou não*)*;*

• *Modelos de predição de clientes inadimplentes;*

• *Diagnóstico de doenças;*

A função logística (também conhecida como sigmoide) só retorna valores entre 0 e 1.

• Os valores de saída da função são interpretados como a probabilidade

da classe ser 0 ou 1;

• Após treinamento do modelo usando regressão logística testamos o

mesmo em um conjunto de dados de teste;

• A principal forma de avalia-lo em DS é por meio da M**atriz de**

**Confusão** para obtenção das métricas de classificações.

**Matriz de Confusão**

**Métricas de validação**

Acurácia = Verdadeiro Positivo VP Total

+ Verdadeiro Negativo (VN)

Precisão = Verdadeiro Verdadeiro Positivo VP Positivo + Falso VP

Positivo (FP)

Recall ou Revocação = Verdadeiro Verdadeiro Positivo VP

Positivo VP + Falso Negativo (FP)

F1 Score = 2 Precisão ∗ Precisão + ∗ Recall

Recall

**Valor Predito**

**l**

**aeRr olaVNegativo Positivo o**

**vita**Verdadeiro Negativo (VN) Falso Negativo (FN)

**geNo**

**viti**Falso Positivo (FP) Verdadeiro Positivo (VP)

**soP**

***Vantagens do modelo Logístico***

• Facilidade para lidar com variáveis categóricas (Resposta ou preditoras)

• Forte resultados em termos de probabilidade

• Facilidade de classificação de instâncias (indivíduos) em categorias

• Requer pequeno número de suposições

• Alto grau de confiabilidade

***Perguntas e serem respondidas***

• O modelo faz sentido?

• O modelo é útil para o objetivo pretendido? Se, por exemplo, o custo da coleta dos dados de uma variável é exorbitante e impossível de ser obtido, isso resultará em um modelo sem utilidade;

• Todos os coeficientes são razoáveis, ou seja, trazem valores que fazem sentido em termo de análise?

• A adequabilidade do modelo é satisfatória? Tem boa Precisão, Recall e F1 Score?

***Interpretação dos Coeficientes***

**PARA CASA: Texto explicativo** Entregar na próxima aula 19/02 Dica: revisar Cap.4, seção 4.3

***META***

• Participação em uma competição de ***Machine Learning*** no *Kaggle*

• Aplicar uma das técnicas de predição utilizadas no curso

• Escrever um artigo sobre modelagem preditiva no *Linkedin*

• *Causalidade*

• *Risco*

• *Identificação*

• *etc*

***Exercício***

**Kaggle**

**Titanic:** Machine Learning from Disaster

***Exercício***

**Variável Definição Chave**

survival Sobrevivência 0 = não,

1 = sim

pclass Classe de bilhetes

**Jupyter Notebook**

1 = primeiro, 2 = segundo, 3 = terceiro

sex Sexo

Age Idade em anos

sibsp Número de irmãos / cônjuges a bordo do Titanic

parch Número de pais / filhos a bordo do Titanic

ticket Número do bilhete

fare Tarifa de passageiro

cabin Número da cabine

embarked Porto de embarcação

C = Cherbourg, Q = Queenstown, S = Southampton

***Exercício***

**Jupyter Notebook**