INF-0615 – Aprendizado de Máquina Supervisionado I

Exercício 3, 4 e 5 - Regressão Logística

Predição do Nível de Colesterol

1 Descrição do Problema

O colesterol é um composto orgânico vital para o bom funcionamento do organismo humano. Ele contribui na produção de hormônios, vitaminas, ácidos envolvidos na digestão e na regeneração celular. Existem dois tipos de lipoproteínas no corpo humano que controlam o colesterol: o HDL (*High Density Lipoprotein*) que retira o colesterol dos vasos sanquíneos e o elimina na urina; e o LDL (*Low Density Lipoprotein*) que retira o colesterol do fígado e o acumula nos vasos sanquíneos. Assim temos o HDL-colesterol (conhecido como colesterol "bom") e o LDL-colesterol (conhecido como colesterol "ruim"). Quanto maior o LDL-colesterol e menor o HDL-colesterol, maior a chance de doenças cardiovasculares[1].

Nesse exercício, iremos predizer se uma pessoa apresenta LDL-colesterol (colesterol "ruim") alto ou baixo baseado em outros compostos presentes no sangue. Para isso, tomamos uma base de dados [2] reportando a concentração de alguns componentes, os quais são descritos abaixo:

- **LBXTR**: Concetração de Triglicerídeos no sangue (em mg/dL).
- **LBDHDD**: Concentração direta de HDL-colesterol (em mg/dL).
- **LBDBANO**: Número de Basófilos (em 1000 células por microlitro).
- **LBDEONO**: Número de Eosinófilos (em 1000 células por microlitro).
- **LBDLYMNO**: Número de linfócitos (em 1000 células por microlitro).
- **LBDMONO**: Número de monócitos (em 1000 células por microlitro).
- **LBXMC**: Concentração média de hemoglobinas (em g/dL).
- **LBDB12**: Concentração de vitaminas B12 (em pg/mL).
- **LBDBCDSI**: Concentração de Cádmio (em umol/L).
- **LBDBMNSI**: Concentração de Manganês (em umol/L).
- **LBXGLT**: Concentração de Glicose nas últimas duas horas (em mg/dL).
- **LBXAPB**: Concentração total de Apoliproteínas (em mg/dL).
- **LBDLDL (class)**: Concentração de LDL-colesterol no sangue. Ele foi categorizado de forma que valores a cima de 130 mg/dL sejam considerados alto (rótulo 1), e valores a baixo desse limiar sejam considerados baixos (rótulo 0). **Esse é o valor alvo que devemos predizer**.

Os valores originais são encontrados no arquivo *labs.csv*. Os valores nos arquivos de treino, validação e teste são aqueles já processados no código (*pre_processing.R*).

2 Atividades

Neste exercício, nós iremos:

1. Inspecionar os dados. Quantos exemplos há de cada classe em cada conjunto? Qual o intervalo de cada feature?
2. Normalizar os dados para que fiquem mais bem preparados para o treinamento.
3. Treinar uma regressão logística para classificar a concentração de colesterol.
4. Classificar os dados de teste.
5. Calcule a matriz de confusão, acurácia, curva ROC, taxa de verdadeiros positivos e de verdadeiros negativos para o conjunto de validação.
6. Explorar aumento de complexidade do modelo por meio de combinação de features e modelos polinomiais. Plotar a curva viés/variância para os dados de treinamento e de validação. Identificar regiões de *underfitting*, *ponto ótimo* e *overfitting*.
7. Explorar técnicas para lidar com desbalanceamento.
8. Explorar técnicas de regularização. Plotar a curva viés/variância com os diferentes valores de regularização. Identificar novamente as regiões de *underfitting*, *ponto ótimo* e *overfitting*.
9. Tomar os melhores modelos dos itens anterioes e classificar os dados de teste. A performance no teste segue a mesma performance na validação ?

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *labs.csv*: base de dados original do problema. Após processamento, os conjuntos abaixo são gerados.
- *cholesterol_training_set.csv*: dados de treinamento;
- *cholesterol_validation_set.csv*: dados de validação;
- *cholesterol_test_set.csv*: dados de teste;
- *pre_processing.R*: código utilizado para pré-processamento da base de dados;
- *support_functions.R*: código com funções de apoio aos três laboratórios;
- *Ex03.R*: exercício 03 para treinar regressão logística com técnicas de aumento de complexidade (polinomial e combinação de feaures);
- *Ex04.R*: exercício 04 para treinar regressão logística regularizada;
- *Ex05.R*: exercício 05 para treinar regressão logística com diferentes técnicas de balanceamento;

4 Referências

1. *Fator de risco: Colesterol*. `https://www.einstein.br/especialidades/cardiologia/doencas-sintomas/colesterol`
2. *National Health and Nutrition Examination Survey*. `https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey?select=labs.csv`