Módulo 3 – Aprendizado de Máquina Supervisionado I

Exercício 3, 4 e 5 - Regressão Logística

Breast Cancer Wisconsin Dataset

1 Descrição do Problema

O câncer de mama é uma doença causada pela multiplicação anormal de células da mama e afeta, na maioria dos casos, as mulheres. Segundo o Instituto Nacional de Câncer (INCA) [1], ele é o segundo tipo mais comum de câncer no Brasil e no mundo (ficando atrás apenas dos tumores de pele do tipo não melanoma) e, para 2018, estão previstos mais de 59 mil novos casos da doença no nosso país.

O objetivo neste exercício é classificar se um tumor é benigno ou maligno, através da análise de algumas características de suas células. Utilizaremos um dataset coletado pelo *University of Wisconsin Hospitals* [2], que contém um total de 699 amostras de células do tecido mamário de diversos pacientes. O dataset foi previamente dividido em conjuntos de treinamento e teste e cada amostra contém as seguintes informações:

- **ID**: um código para identificar a amostra;
- **Clump Thickness**: células de tumores benignos tendem a se agrupar em uma única camada, enquanto células cancerosas se agrupam em múltiplas;
- **Uniformity of Cell Size**: medida da variação média do tamanho das células. Uma vez que células cancerosas tendem a ter uma variação grande em tamanho e forma, essa medida pode ser importante para classificar um tumor;
- **Uniformity of Cell Shape**: semelhante ao anterior, porém referente à forma da célula;
- **Single Epithelial Cell Size**: relacionado com a uniformidade acima, células epiteliais muito grandes podem ser malignas;
- **Marginal Adhesion**: células normais se agrupam, enquanto que as cancerosas tendem a perder essa habilidade. A perca de aderência é um sinal de um tumor maligno;
- **Bare Nuclei**: termo utilizado para células cujo núcleo que não está envolto em citoplasma, tipicamente visto em tumores benignos;
- **Bland Chromatin**: caracteriza a textura uniforme encontrada nos núcleos de células benignas;
- **Normal Nucleoli**: os nucléolos são pequenas estruturas presentes no núcleo. Em células normais, os nucléolos são muito pequenos, enquanto que em células cancerosas eles podem ser maiores e mais numerosos;
- **Mitoses**: uma medida da velocidade de multiplicação das células tumorais;
- **Class**: classificação do tumor ("2" para benigno e "4" para maligno);

2 Tarefas

Neste exercício, nós iremos:

1. Inspecionar os dados de treinamento. Quantos exemplos há de cada classe? Qual o intervalo de cada feature?
2. Treinar uma regressão logística para classificar o tumor.
3. Classificar os dados de teste.
4. Calcule a matrix de confusão, acurácia, curva ROC, taxa de verdadeiros positivos e de verdadeiros negativos para o conjunto de teste.
5. Explorar técnicas para lidar com desbalanceamento.
6. Explorar técnicas de regularização.

3 Arquivos

Os arquivos disponíveis no Moodle são:

- *breastCancer_train.data*: dados de treinamento que serão utilizados no exercício;
- *breastCancer_test.data*: dados de teste que serão utilizados no exercício;
- *breastCancer_train_unbalanced.data*: dados de treinamento para explorarmos técnicas de balanceamento de dados;
- *breastCancer_test_unbalanced.data*: dados de teste para explorarmos técnicas de balanceamento de dados;
- *breastCancer_train_regularization.data*: dados de treinamento para explorarmos a influência da regularização;
- *breastCancer_test_regularization.data*: dados de teste para explorarmos a influência da regularização;

4 Referências

1. *Câncer de Mama*. Instituto Nacional de Câncer José Alencar Gomes da Silva. `http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/mama`
2. *Breast Cancer Wisconsin dataset*. UCI Machine Learning Repository. `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)`.