Trabalho 2

Implemente 2 dos 4 algoritmos vistos em sala de aula para o processo de classificação (KNN, RNA, SVM ou Árvore de decisão).

O objetivo do trabalho é classificar se a amostra da imagem de 3 células distintas de câncer pode identificar se o câncer é benigno (B) ou maligno (M) para o câncer de mama.

O arquivo cancer_breast.csv possui 357 amostras benignas e 212 malignas. Cada amostra possui as informações extraídas por PDI (Processamento Digital de Imagem) de 3 células distintas de um mesmo tumor que autopsiado.

A primeira coluna do arquivo indica se o tumor é maligno ou não e as 30 colunas a seguir são referentes as seguintes informações de cada uma das 3 células

1. Raio
2. Textura (Desvio padrão em escala de cinza)
3. Perímetro
4. Área
5. Variação do raio da célula
6. Compactação da célula
7. Cavidade
8. Qtde de concavidades da célula
9. Simetria
10. Dimensão fractal (Aproximação fractal do elemento real)

A linguagem de programação a ser utilizada é livre.

É necessário entregar um programa que receba como entrada o arquivo CSV e o divida em 10 blocos, onde em cada um deles um bloco é utilizado para treino e outro para teste. Como no esquema a baixo

Cenário 1:	1-100	101-200	201-300	301-400	401-500
Cenário 2:	1-100	101-200	201-300	301-400	401-500

…

Para cada cenário os conjuntos de blocos em branco foram utilizados para treino enquanto os de cinza foram testados. Ao final, apresente os resultados da quantidade de acertos e erros para cada grupo.

Dicas:

Em python a função genfromtxt do numpy já lê arquivos CSV.

https://docs.scipy.org/doc/numpy-1.14.0/reference/generated/numpy.genfromtxt.html

Trabalho 3

Faça o tratamento da análise do sentimento de tweets em relação às empresas aéreas dos estados unidos referentes a feveriro de 2015. O dataset possui 3 colunas.

1. Sentimento → Positivo ou negativo em relação a empresas
2. Empresa → As empresas citadas
3. Texto → Texto extraído do tweet

Você deve gerar o dataset (de preferência o CSV) com o texto se tornando um conjunto de atributos (neste caso colunas) para cada comentário, retirando apenas aquilo que é importante. Para isso, utilize a biblioteca NLTK.

https://pythonspot.com/category/nltk/

Nesse tutorial tente fazer o processo de “tokenizing”, remoção de “stop words” e aplique o “stemming”. Dessa forma você terá um conjunto de palavras relevantes dentro do dataset, apenas converta-o para csv e utilize os 2 algoritmos de classificação não utilizados no trabalho. Pode ser utilizado tanto o Weka quanto o SKLearn ou outra biblioteca similar para o processo de predição.

Data de entrega: 16/12/2018 às 19hs.