Trabalho 5: Detecção de Spam com Aprendizagem de Máquina

Introdução

O objetivo deste trabalho é utilizar a ferramenta scikit-learn para comparar alguns algoritmos de aprendizagem de máquina em um problema de detecção de spam. Os algoritmos a serem comparados são **regressão logística** e **k-NN**. Você pode tomar como ponto de partida o script de exemplo apresentado na aula de 26/Nov/2018.

Você deve preparar um relatório em PDF que deve ser entregue pelo Moodle. O resultado deste trabalho é um relatório em formato PDF detalhando suas atividades, além de um script em Python, com os comandos para realização do que se pede.

A base de dados a ser utilizada nesse trabalho é conhecida como Spambase e foi produzida por um grupo de pesquisa do Hewlett-Packard Labs. Ela está disponível publicamente no UCI Machine Learning Repository e contém 4601 mensagens de e-mail (sendo que 1813 são spam e 2788 não são spam). Cada mensagem é representada por 57 atributos numéricos sendo que a maioria desses atributos representa a frequência de uma determinada palavra na mensagem.

A base de dados é fornecida no formato ARFF. Para ler o arquivo com Python, use os comandos a seguir:

```
import arff, numpy as np
dataset = arff.load(open('spambase.arff', 'rb'))
data = np.array(dataset['data'])
```

Após ler o conjunto de dados, faça a divisão em conjunto de treinamento e conjunto de testes. Reserve 20% do total para o conjunto de testes. Você irá

usar esse conjunto de testes para medir a precisão dos modelos preditivos que você irá gerar, e apenas para isso.

Após a primeira divisão dos dados, faça uma segunda divisão, dessa vez apenas sobre os 80% dos dados não pertencentes ao conjunto de testes. Chamemos esse conjunto de dados remanescente de D. Essa segunda divisão deve gerar o *conjunto de treinamento* propriamente e o *conjunto de validação*. Reserve 20% de D para o conjunto de validação.

Você agora deve usar o conjunto de validação para determinar o melhor valor de k para o algoritmo k-NN. Para isso, faça o seguinte: defina uma faixa de valores de k (semelhante ao código de exemplo fornecido em aula). Em seguida, para cada valor de k, gere um modelo preditivo usando apenas o conjunto de treinamento. Em seguida meça a precisão (acurácia) desse modelo usando apenas o conjunto de validação. O valor de k que você deve selecionar deve ser aquele que produzir a maior precisão sobre o conjunto de validação.

Agora que você determinou o melhor valor de k para o kNN, você deve executar tanto o kNN quanto a regressão logística sobre o conjunto de dados D. Com isso, você irá gerar dois modelos preditivos. Finalmente, use o conjunto de teste para estimar o poder preditivo (acurácia) de cada um desses dois modelos. Qual dois dos modelos se saiu melhor? Reporte os valores de precisão em seu relatório.