Desafio 1 - Classificador de ondas

Gabriel Mascarenhas Ribeiro de Paula

Neste exercício, é apresentada uma tabela em csv em que suas colunas 22 são separadas pelo caractere de vírgula e as 5000 linhas por quebras-de-linha. Tal tabela corresponde à um conjunto de formas de ondas, nos quais as 21 primeiras colunas são os atributos das ondas e a última coluna há indicação de qual das três classes de onda existentes ela pertence.

Para carregar a tabela em *python* utilizou-se de um leitor de csv, repassando seu conteúdo a uma matriz. Após isso são separados os dados que servirão para a montagem a árvore de classificação a ser utilizada. Assim, como entradas são utilizados os atributos das formas de onda e, como saída um *array* contendo as classes atribuídas a cada uma.

Os dados de entrada e de saída correspondentes foram divididos em duas classes, de treinamento (70%) e de teste (30%). Tal divisão é imprescindível para a avaliação da qualidade da classificação, uma vez que um modelo só pode ser considerado válido se for testado com um conjunto de dados inéditos ao classificador. Métodos de validação mais confiáveis não foram utilizados nesse projeto, uma vez que não foi enunciado um critério de rigor quanto à qualidade da classificação. Usou-se um medidor de acurácia apenas para verificar se a classificação estava no caminho correto.

Criou-se uma árvore de classificação com os dados de treinamento utilizando o pacote de ferramentas *scikit-learn*. Impôs-se, como restrição para sua construção, o limite de dez folhas, ou seja são escolhidas dez regras de classificação para a determinação das classes.

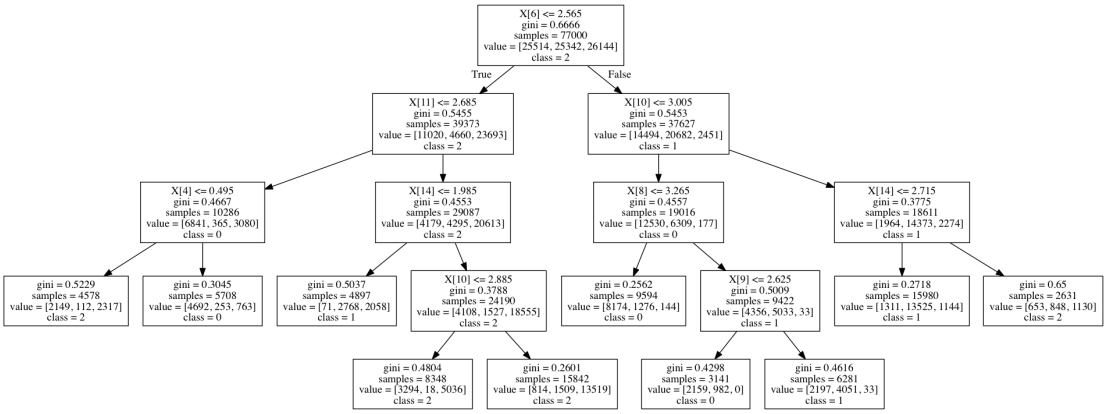

Figure 1: Árvore de Classificação construída

A árvore construída, pode ser representada no formato de um grafo orientado em arquivo ".dot" (veja o arquivo tree.dot no diretório raiz). Tal representação é insumo para a construção da árvore em forma de imagem no formato png (Figura 1), através do aplicativo *graphviz*.

A representação da árvore nos trás informações importantes, tais como: o atributo utilizado como critério para as divisões; o valor do índice de *gini*, que indica o nível de desigualdade das classes quando segundo os valores do atributo analisado; o número total de amostras analisadas por trecho e qual a divisão de elementos de cada classe está presente em cada nó. Quanto maior o número de itens de uma classe nas folhas em detrimento do número de itens de outras classes, maior a qualidade da classificação.

Por fim, as regras que constituem a árvore e definem a classificação estão dispostas abaixo:

Table 1: Regras de Classificação

Regras para classificação da Classe 0 :
SE X[6] <= 2.565 E X[11] <= 2.685 E X[4] > 0.495 OU
SE X[6] > 2.565 E X[10] <= 3.005 E X[8] <= 3.265 OU
SE X[6] > 2.565 E X[10] <= 3.005 E X[8] > 3.265 E X[9] <= 2.625

Regras para classificação da Classe 1
SE X[6] <= 2.565 E X[11] > 2.685 E X[14] <= 1.985 OU
SE X[6] > 2.565 E X[10] <= 3.005 E X[8] > 3.265 E X[9] > 2.625
OU
SE X[6] > 2.565 E X[10] > 3.005 E X[14] <= 2.715

Regras para classificação da Classe 2
SE X[6] <= 2.565 E X[11] <= 2.685 E X[4] <= 0.495 OU
SE X[6] <= 2.565 E X[11] > 2.685 E X[14] > 1.985 E X[10] <= 2.885 OU
SE X[6] <= 2.565 E X[11] > 2.685 E X[14] > 1.985 E X[10] > 2.885 OU
SE X[6] > 2.565 E X[10] > 3.005 E X[14] > 2.715

Links importantes:

http://www.graphviz.org/
http://scikit-learn.org/stable/