Assignment: Ensemble Methods

DSA 8401: Applied Machine Learning

Master in Data Science and Analytics

Strathmore University
@iLabAfrica Centre

1 Goals

This week's activity focuses on studying of a new group of methods that works very well as an starting point in most of machine learning projects. The specific goals are:

- Review the pre-processing steps that every project must follow.
- Try two methods: Random Forest (RF) and Gradient Boosting Trees
- Study the influence of different parameters in the correct generalization of a model

2 Assignment Description

In this activity we will work with the dataset called "Fetal Health Classification", which you can found at https://www.kaggle.com/andrewmvd/fetal-health-classification.

It is requested:

done 1. Import the dataset using Pandas. Does it have missings?

done 2. Our objective will be to distinguish subjects with normal fetal status from the rest. What should we do in the data set?

review 3. Now analyze the data and do the pre-processing that you consider appropriate. It is requested here to analyze the data incorporating medical knowledge (in case you have) to detect possible outliers; possible incoherent values, etc. Consider the use of seaborn's pairplot function (sns.pairplot()) to tackle groups of variables in one go (although probably not all the dataset at once).

done 4. Is this an unbalanced problem?

done 5. Apply normalization on the dataframe if necessary and divide the available examples into 80% for training and 20% for testing, randomly. What code have you applied?

done 6. Train a Random Forest (RF) model by adjusting the number of trees to 100, 300 and 1000. Use 5-fold CV and choose the number of trees that obtains the best balanced accuracy.

done 7. Comparing the results obtained with the three models, does it fit with what you expected to obtain?

done 8. Now evaluate the chosen model on the test data. What metric value do you get? Does the model generalize well?

done 9. We will now train a Gradient Boosting Trees model using 5-fold CV. We will consider the default parameters and we will only vary the number of estimators (100 and 200) and the learning rate (0.01, 0.1 and 1). Which model obtains better balanced accuracy?

done 10. Analyze how the results obtained are modified depending on the two parameters tested. Does it fit with what should be expected?

done 11. With the best set of parameters selected, evaluate the model on the test data. Does it generalize well in this case?

done 12. Finally, obtain both the ROC curve of the model and the area under the curve. At what point of the ROC curve would you work if you had to apply it in practice? Justify your answer.

3 Hand in

The students will upload in the eLearning platform the Jupyter notebook. The notebook will contain the python code, the plots and especially personal responses to the previous questions. If there is no personal comment the mark will be F. Marks C,B and A will depend on the variety and quality of the personal comments.