Assigment: Regularization and SVM

DSA 8401: Applied Machine Learning

Master in Data Science and Analytics

Strathmore University

@iLabAfrica Centre

1 Goals

This week's activity focuses on studying the importance of regularization methods and practising with a new way of modelling data: Support Vectors Machines (SVM). The specific goals are:

- Practice the preprocessing of the data to get a suitable dataset.
- Study the generalization of the model when we use regularization techniques.
- Use and analyse SVM

2 Assignment Description

In this activity we will work with the dataset called "Wisconsin Diagnostic Breast Cancer (WDBC)", which you can find at `https://www.kaggle.com/uciml/breast-cancer-wisconsin-data`.

It is requested:

done 1. Import the dataset using Pandas. Do you have missing values? If so, what option have you adopted to resolve it?

done 2. Now analyze the data and do the pre-processing that you consider appropriate. Here we are asked to analyze the data incorporating medical knowledge to detect possible outliers, possible incoherent values, etc. You can use seaborn's pairplot function if you see fit (sns.pairplot()) to tackle groups of variables in one go (although probably not the entire dataset at once).

done 3. Just by visualizing the point clouds of the variables (you don't have to calculate), do you see a correlation between variables? Indicate the pairs of variables in question.

done 4. Now look at the variable 'id'. Is it convenient to use this variable? If you consider taking any action, indicate which one.

done 5. Now analyze the 'diagnosis' variable, which would be the object of our study (predicting whether it is a benign or malignant case). What's the problem with putting it in any algorithm? What solution are you going to adopt?

done 6. Is this an unbalanced problem?

done 7. Apply normalization to the dataframe and divide the available examples into 80% for training and 20% for testing, randomly.

done

8. Now apply standard logistic regression (without regularization) to obtain a first data on the accuracy of the model. What parameters have you chosen increating the logistic regression model?

9. Compare model accuracy on test and training data. Does the model generalize well?

10. Now we will use L1 regularization to try to simplify the number of variables without losing performance. What variables have you been able to eliminate and what is the performance of the algorithm on both training and test data? What regularization value have you chosen?

11. What variables has the model chosen? Does it make sense from a medical point of view? (Even than you are not physicians try to guess).

12. Train an SVM model with Gaussian kernel and compare results. For both the gamma parameter and C we will try the values 0.1, 1 and 10 and we will work with 5-fold CV. What combination of values provides the best result? What is the accuracy of the model?

13. Finally, choose the best model obtained, visualize its confusion matrix and analyze the result with what was seen in week 3.

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.