Golden Gate University

MSBA 326 Predictive Analytics & Machine Learning

Instructor: Jahan Ghofraniha

Homework on Decision Trees using scikit-learn

Reading Assignment:

1. Review Decision tree slides
2. What are the precision and recall definitions? Explain why there is a tradeoff between the two? You can use graphs or any other tools to answer this question.
3. What is the definition of F1 score and how do you interpret a high F1 score?
4. Use the baseball salary dataset and the exploratory data analysis scheme to determine visually which are the candidate features for the model (correlation and scatter plot)
5. Calculate the descriptive statistics for the dataset. What is this information good for?
6. Plot the histogram of each column. How do you interpret the distribution plots?
7. Use the log(salary) as your output and pick six features as input for your data (use the exploratory analysis as a basis for the choice of input features).
8. Develop a regression decision tree model for the dataset base on default setting of the regressor,(i.e. use DecisionTreeRegressor() without any input. Check out the documentation for DecisionTreeRegressor()). You can save the image of the decision tree by right clicking on the console and save image/copy image and paste it into a paint or any other graphic package and save it as a .png file so you can look at it.
9. How many levels does the default decision tree have based on the six features (tree depth)?
10. Run the system again with a depth of 2 and compare the performance measures in the two cases. Explain the difference in the context of performance measure as it relates to variance/bias trade off.
11. Try the code a third time with a depth of 3 and make the same comparison as in step 10.
12. If you were to optimize this system using the tree depth (max_depth) as the hyper-parameter what would be your suggestion?
13. Use the bank note authentication dataset as given in the sample code to build a classification tree.
14. Use the sample code given on page 89 to plot the recall/precision curve.
15. Based on step 14 choose a suitable threshold to achieve a high precision. Would this strategy work in every case? It seems with a high enough threshold you can achieve a good precision, is this method a good approach to optimizing your classifier?

16. Plot the ROC curve (use sample code given on page 91) and use the roc_auc_score function (sample code given on page 92 of the textbook) to calculate ROC score.
17. (Optional) Try setting the tree depth (max_depth) to 2 or 3 and repeat steps 13-16 and compare the performance measures.
18. Include your code, the results and explanation of the results either as a Jupyter file (.ipynb) or as comments in your code. Include your plots as screen shots/pictures and upload it to Moodle.