Golden Gate University

MSBA-326, Fall 2018

Instructor: Jahan Ghofraniha

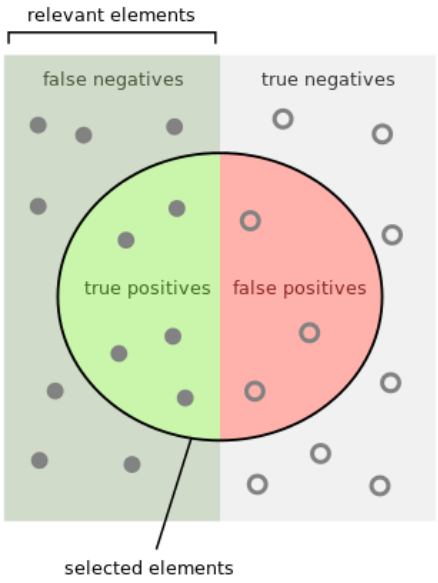
Homework 5 Solution Decision Trees using scikit-learn

Reading Assignment:

1. Read chapters 3 and 6 from the textbook

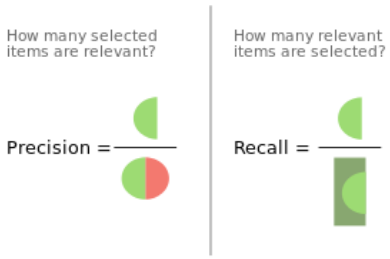
NA

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.



$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$



If the threshold on false positive is increased, i.e. we get stricter on positive identification it becomes easier to miss a true positive so we tend to miss a lot more and the number of false negatives goes up. When we make less type I error the tendency for type II error goes up. That is the reason for tradeoff between the two.

3. What is the definition of F1 score and how do you interpret a high F1 score?

F1 score is the harmonic mean between precision and recall. It is a single number that combines the two measures. Since the impact of both measures are considered in one number or score, it is better than individual measures such precision and recall alone as one can have a 100% precision while a very bad recall.

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)

We should consider two criteria for determination of dominant features. The first is a high correlation of one feature with the output of the model. In this case, we should eliminate one of the correlated features or use a model reduction technique such as PCA. The second criterion is the lack of strong correlation between the features (input elements/predictors).

5. Calculate the descriptive statistics for the dataset. What is this information good for?

Descriptive statistics are the location measures (mean, variance and percentiles) of the distribution curve for each of the features. They are numerical representation of the shape of the distribution curve for each feature. We can use the descriptive statistics to determine if we should use data transformation techniques such as removing the mean /mean centering and standardization. The general shape of the histogram/PDF of each feature will also tell us if we need nonlinear transformation such log on input or output data before using the data in our modeling effort.

6. Plot the histogram of each column. How do you interpret the distribution plots?

We should look at the shape of the histogram and see how close they are with a Gaussian like distribution. Skewness (third moment) and peakness (Kurtosis) of the data/features are also important to consider. The normal approach to deal with skewness is log transform or Box-Cox transform (natural log or normalized exponentiation). Kurtosis shows how far a distribution is from a uniform transform. In regression models a large skewness can result in a large or wrong intercept in a linear regression model.

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).

The proper way of choosing the most important features is to use the over-fitted tree since decision trees by default implement the feature importance when a feature is picked for splitting. If you look at the fully grown tree, the top six level features are (sorted by the most important first):

X7 = CAtBat

X1 = Hits

X10 = CRuns

X4 = RBI

X0 = AtBat

X9 = CHmRun

It is important to perform step 8 before finding a proper answer for step 7.

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.

Check out the distributed code.

9. How many levels does the default decision tree have based on the six features (tree depth)?

15

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.

A depth 2 tree will have less variance and the MSE will be larger. However, it will have more bias and less variance. The key idea here is to figure out if depth 2 is under-fitted. The best way to determine what is the correct setting of a hyper parameter is to use cross validation to compare models and pick the right one.

11. Try the code a third time with a depth of 3 and make the same comparison as in step 10.

If we were to compare depth= 3 to depth=2, we need to compare them with respect to the MSE error in the test set. The variance of the Depth 3 tree will be larger than depth = 2 while the bias will be smaller. The correct determination of the right depth is obtained with cross validation.

12. If you were to optimize this system using the tree depth (max_depth) as the hyper-parameter what would be your suggestion?

The proper and sure way of determining the correct max_depth is to use cross validation. The manual approach in the absence of cross validation is to change max_depth and monitor the MSE of the test sets where the MSE settles to a steady value pick the first value that results in the settled MSE trend.

13. Use the bank note authentication dataset as given in the sample code to build a classification tree.

Check out the code.

14. Use the sample code given on page 89 to plot the recall/precision curve.

Check out the code.

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?

Generally we want to achieve a good balance with a good precision and recall. This normally happens by plotting the precision and recall vs. the threshold value. The intersection of these two curves is a good starting point for the choice of threshold. Depending on the application you may want to increase the precision threshold or lower it from the intersection of the two curves.

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.

Check out the code.

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 SMCMBA dropbox.