DAT16 SF: HOMEWORK 5 ASSIGNMENT

Assigned: Wednesday, August 26, 2015
Due: Monday, August 31, 2015, before class
Review Due: Wednesday, September 2, 2015, before class

The purpose of this homework is to review what we've learned about Support Vector Machines, Decision Trees and learning curves.

HOMEWORK QUESTIONS

DUE MONDAY:

1. Use the provided dataset cancer_uci_HW5.csv
   (which comes from:
   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)
)
2. Import the dataset and check for balance between the two classes. If the ratio is less than 60/40 rebalance the classes to 50/50 using one of the strategies learng in class.
3. Are the features normalized? If not, use scikit-learn standard scaler to normalize them.
4. Train a linear SVM, using CV accuracy as score (use the scikit learn function, not the one used in class).
5. Display the confusion matrix and calculate Precision, Recall, ROC curve ad AUC.
6. Do you need more data or a better model? Plot a learning curve to find out.
7. Repeat steps 3-4-5 using a Decision Tree model. Are results better or worse?
8. How could you combine the two models?

BONUS POINTS:

Try using an SVM with RBF kernel. Is it better?

DUE WEDNESDAY:

1. Go to your new assigned review-buddy's repo
2. Read through your buddy's ipython notebook and make sure you understand what he/she is doing.
3. Open an issue in his/her repo and write comments on the things you don't understand and on the things you like in his/her code.
4. Quote the instructors in the comments so that we get notified about the open issue