

DAT16 SF: HOMEWORK 3 ASSIGNMENT

Assigned: Wednesday, August 12, 2015

Due: Monday, August 17, 2015, before class

Review Due: Wednesday, August 19, 2015, before class

The purpose of this homework is to review what we've learned about classification problems, cross-validation, KNN, Naïve Bayes

HOMEWORK QUESTIONS

DUE MONDAY:

1. Download the Pima Indians dataset from UCI here:
<https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>
2. Describe the content of the dataset in your own words
3. Describe the features and formulate hypothesis on which may be relevant in predicting diabetes
4. Import the dataset to a Pandas dataframe and explore the data:
 - a. Are there any missing data or NULL values? How could they be imputed? Make a choice and impute them or drop them. Justify the choice.
 - b. How many features are there? Are they normalized?
5. Use the KNN classifier from Scikit-learn to predict diabetes occurrence
6. Use Scikit-learn cross-validation routine to evaluate the accuracy of your model with a 5-fold CV.
7. Plot the 5-fold CV accuracy score as a function of K for k up to 50 neighbors
8. Use the Naïve Bayes classifier from Scikit-learn to predict diabetes occurrence
9. Compare the 5-fold CV score for Naïve Bayes and for KNN to find which model is more accurate

BONUS POINTS:

Read through this blog post for an implementation from scratch:

<http://machinelearningmastery.com/naive-bayes-classifier-scratch-python/>

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