DAT10 SF: HOMEWORK 7 ASSIGNMENT

Assigned: Friday, November 21th
Due: Monday, November 24th, midnight
Review due: Wednesday, November 26th.

The purpose of this homework is to gain deeper understanding of Support Vector Machines and grid search.

DATA & CONTEXT

For this assignment we will use the Wine dataset that you can find here:

https://archive.ics.uci.edu/ml/datasets/Wine

It contains 13 chemical measurements on 178 wines from 3 regions of Italy. The features are named (x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13) and the labels are the regions.

HOMEWORK QUESTIONS

1. Classify the raw data using a linear SVM. Do you need to perform several binary classifications or does scikit-learn support multi-class classification with SVMs?
2. Cross validate the result
3. Preprocess the data with a normalization step, using the tools explained here: http://scikit-learn.org/stable/modules/preprocessing.html#standardization-or-mean-removal-and-variance-scaling
4. Repeat the classification performed in step 1 using a linear SVM and crossvalidate the result. Is it better or worse?
5. Learn about pipelines here:
   http://scikit-learn.org/stable/modules/pipeline.html
   implement a pipeline that comprises:
   - a preprocessing step
   - a classification step

   and run the pipeline on the raw data (not normalized)
6. Try varying the value of C or the type of kernel. Do you get better results?
7. Learn about grid search here:
   http://scikit-learn.org/stable/modules/grid_search.html
   and feed your pipeline classifier to the grid search. Explore a range of values for C, gamma and the type of kernel. Can you find an optimum value?