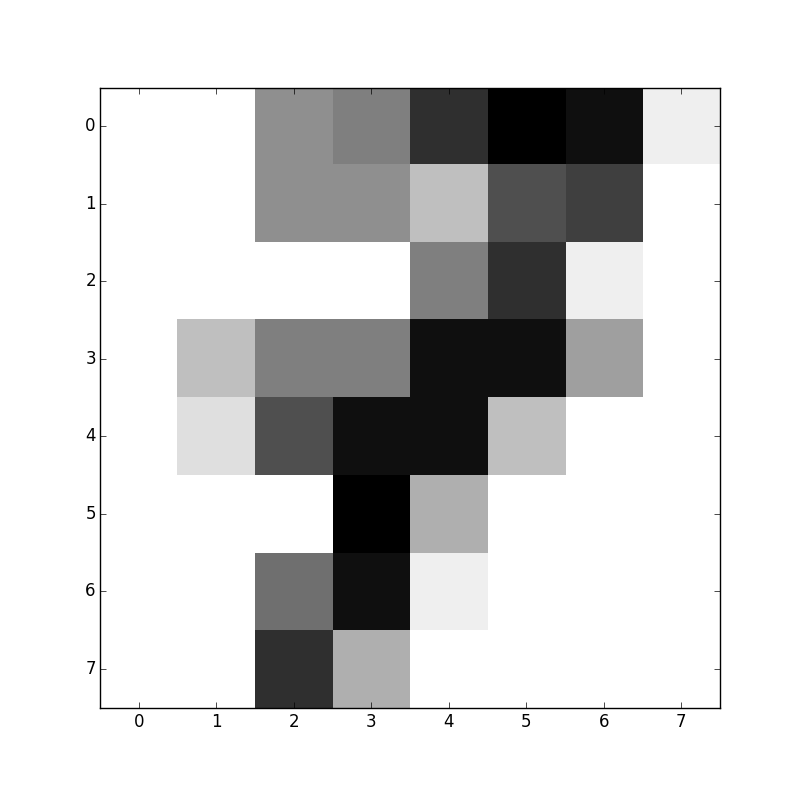
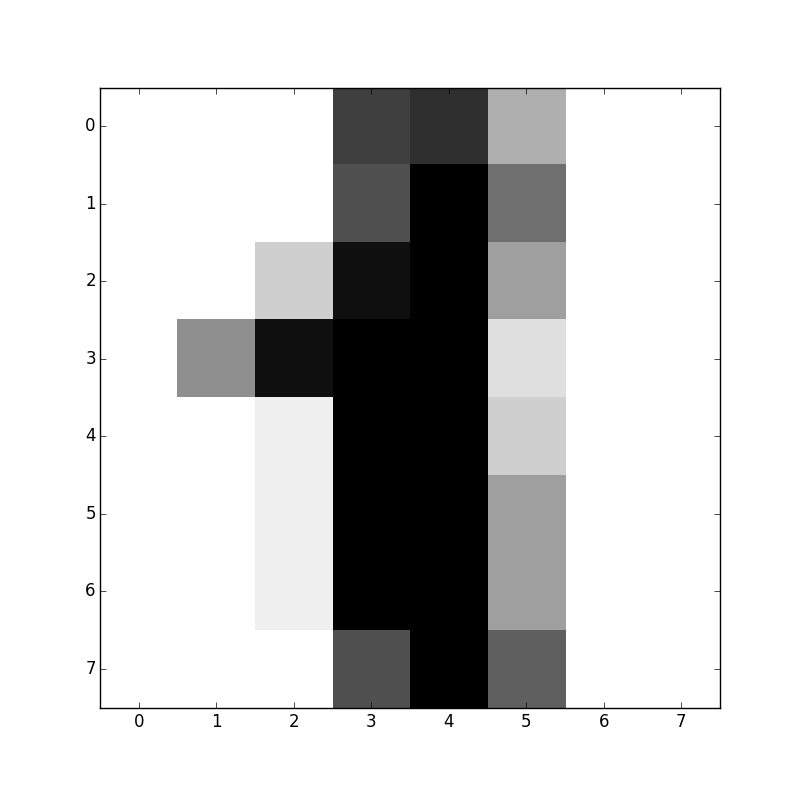
INF-552 [201701-32402] Machine Learning for Data Informatics

Programming Assignment 1

**Due date:** Sunday, February 26th, 11:59pm PT. Late submissions will be hit with at least 20% penalty.

The purpose of this homework is to give you hands-on experience with the process of training, validating and testing classifiers. To simplify the homework, we will be using a toy dataset in scikit learn. [The Digit Dataset](http://scikit-learn.org/stable/auto_examples/datasets/plot_digits_last_image.html#sphx-glr-auto-examples-datasets-plot-digits-last-image-py) consists of 8x8 images of hand written digits (0~9). In other words, each data point is an 8x8 matrix with an integer label. Meanwhile, each element in a matrix is also an integer in the range of 0~16 representing the grey level of a pixel.



Q1. Calculate the statistic metrics (i.e. mean, standard deviation) of the first grey-level image in the dataset. (1 point)

Q2. Calculate the mean value of each prime number digit (2, 3, 5, 7). Then, use [boxplot](http://matplotlib.org/api/pyplot_api.html#matplotlib.pyplot.boxplot) to show and compare the distributions of those four categories. (3 points)

Q3. Binary classification task: Is it a prime number? (16 points)

As a first step, you are asked to assign new labels (integer ‘1’ as TRUE; ‘0’ as FALSE) for all digits. Flatten images by reshaping feature matrices (8x8) to vectors (1x64). You can also include the statistic metrics in the feature vector and apply other feature engineering techniques. Write one paragraph to describe your data processing approach.

For experiment A, use SVM classifier with linear kernel and conduct 5-fold splitting manually. Consider stratified strategies we’ve mentioned in the tutorial class to deal with data skewness. Hold out 1 fold as the testing set. Report the sample size for both categories in the training set. Execute the 4-fold cross-validation strategy manually on the remaining 4 folds (3 for training set and 1 for validation set). Validate the hyper-parameter ***C*** with values on the log-space from 0.01 to 100. Make a chart to report the best ***C*** and the average training, validating and testing accuracies for each of the 4 validation settings. Write one paragraph to describe your approach. Use evaluation metrics such as accuracy, confusion matrix, F1 score and plots to show your results.

For experiment B, compare the linear SVM classifier with two other classifiers for 4-fold cross-validation. Just keep in mind that always use exactly the same data splits for the testing process in order to assure that the results are comparable. For each classifier, you are asked use [GridSearchCV](http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html) and tune one hyper parameter (if there is any) with at least 3 values. Report the performance of each model with the best hyper parameter. Use evaluation metrics and plots to show your results and thoughts within one page.

Please make sure to submit your source code alongside your report. Have fun!