MATH 401, Midterm 2, FALL 2015

Report Due: November 20, 2015

1) Download the dataset of handwritten digits collected by USPS and divide them into two sets of 5500 images: the training set and the testing set. The goal of this project is to develop and test methods for classification of the handwritten data, which are optimized on a new representation of the dataset and then applied to the testing set to assess the performance of the developed methodology.

(With the prmission of the instructor, you can use some other data set, of comparble size and difficulty to the handwriten digits data.)

2) Find and describe a method to represent the digit images (or elements from your dataset) as vectors.

3) Use the nearest neighbors classification scheme, or any other suitable classifier that you may already know, to verify the success rate of your classifications applied to original vectorized data. Optimize the parameters (inlcuing but the selected metric) to maximize the global success rate.

4) Next, use a data representation based on eigenvectors of Graph Laplacian for this dataset, to find the best possible representation of the handwritten digits dataset, for the purposes of classification and identification of those digits. The project work includes weight and neighborhood selection, parameter optimization for Graph Laplacian representation, as well optimization of digit classifications.

5) Write a report which details your choice of methods, and provides a comparative analysis of digit classifications based on original data representation and on the eigenvectors of the Laplacian Eigenmaps. Describe the methods used, and provide interpretation of your own results.

Instructions:

1) You are **not** allowed to communicate with anybody about any methods to solve this exam, this includes electronic means of communication, such as e.g., posting questions on forums etc.

2) You may use all existing "passive" sources, such as books, tutorials, instructions, online video, etc.

3) You must make sure that images or other sources of data you are using (in case you want to work with different data than provided) are yours, or are in public domain. You must ensure that all conditions for using such freely available images are met. You may ask someone for permission to use their images, as long as this does not contradict (1).

4) There are no longer any constraints on matlab usage: you may use all available functions or codes - as long as they are in public domain. You must acknowledge all matlab functions you use. If you are using matlab functions which are not part of standard matlab distribution, you must detail their usage and provide theoretical justification.

5) In your paper please include theory, codes, documentation, and numerical and imagery examples.

Grading:

1) The exam is worth 100 points.

2) The project has a theoretical component, worth up to 30 points.

3) The project requires you to implement an algorithm on your own, worth up to 30 points. (This includes explicitly stating the algorithm in pseudo-code and a providing code with proper documentation.)

4) The project requires you to perform a data analysis task, worth up to 40 points. (This includes specifying the quantitative methods of comparison or analysis.)

Hard copy of the exam is due on November 20th, 2015, at the beginning of the class.