Project 3

CSE 474/574

In this assignment, you will use machine learning to solve a problem with a real dataset. You will be using the MNIST dataset. Your task is to classify images of handwritten digits to their corresponding numbers (0 to 9). You will have 60,000 labeled images to train and 10,000 labeled images to test your machine learning model. This is an open assignment where you are not expected to use a specific machine learning algorithm to solve the problem. You are welcome to use any approach.

Description of Dataset

The MNIST dataset is a well known public dataset which consists of handwritten digits. It has a training set which has 60,000 examples and a test set of 10,000 examples. It was created by remixing the samples taken from American Census Bureau employees and American high school students. The training set contains 30,000 images from employees and 30,000 images from students while the test set contains 5,000 images from employees and 5,000 from students.

The dataset can be accessed using the following code snippet

```
from keras.datasets import mnist
(train_X, train_y), (test_X, test_y) = mnist.load_data()
```

Deliverables

1. Your code which takes the labeled data and returns a model that can be used to make predictions for the test data. The code should be submitted in the form of an iPython notebook or a python file. [30%]
2. A write-up in the form of a pdf document that includes the following: [70%]
    a. An introduction describing at a high level some approaches that you considered, and why you considered them.
    b. A description of your submitted solution, including any data processing, algorithms used, etc.
    c. A section describing some empirical results from your solution. That is, some experiments that demonstrate that your solution is sensible. This section must include a comparison of your approach against at least two other approaches that you have tried. Additionally, you may choose to include other experimental results such as an evaluation of a regularization technique to prevent overfitting, a comparison of different model parameters, such as tanh vs. sigmoid neurons/the number of neurons in a neural network, or different kernels/kernel parameter settings in an SVM, Random Forests etc. The idea is

to justify your approach, and to demonstrate the different factors that you considered for your submitted solution.

d. A conclusion summarizing your work and findings. If your method performed poorly on the test data then you may want to include an explanation as to why and suggest how your method may be improved.
e. References to any code, methods, or ideas that you used that are not your own.
f. Any special instructions that are required to run your code.

Grading

Your grade will be based on 3 criterias:

1. Design and analysis of experiments: You should follow the best practices of experimenting. Using cross validation and performing statistical tests to prove the significance of your results will earn you points. Moreover you should follow a logical way to find and select the best hyperparameters.
2. Report organization: You should present your results in a clear and concise fashion. You should create plots and tables to show important trends and results. Your reports should not exceed 6 pages including figures. It can be less than 6 pages, so don't try to make it too long. Reports should be typed, not hand-written. You should give attention to details such as naming figures correctly, using the right mathematical terms, labeling the axes of plots, etc.
3. Comments on method and results: Your method should be motivated and justified from the structure of the dataset. You should explain for which cases your method worked well and for which cases it failed and why. You should comment on your overall performance and explain why it is high/low. You should also try different variations of your method (regularization, different kernels/neuron types/distance metrics etc.) and comment on what changes when you try different variants.

Hints

You are free to try any methods or approaches that you wish for this assignment. There are many possible solutions; in fact this dataset has been used as a benchmark for active research. There are many approaches that you can try, including discriminative methods (k-nearest neighbors, SVMs, kernel SVMs, neural networks, etc.), generative methods (mixtures of Gaussians, etc.) and dimensionality reduction (PCA, etc.). Yoursolution does not have to be limited to simply applying machine learning models. In the real world, the best approaches can often be simple ones that rely on clever data processing

The handwritten digit prediction task is not intended to be a stressful exercise; instead it is a chance for you to experiment, to play and to hopefully have fun! Start with simple methods that work more or less “out of the box” and go from there.