CPTS 437: Introduction to Machine Learning (Spring 2018)

Homework #5

Due 4/17/2018

- You need to submit a report in hard-copy before lecture and your code to Blackboard.
- Hard-copy is due in class before lecture and electronic copy is due 2:50PM on Blackboard on the due date.
- Unlimited number of submissions are allowed on Blackboard and the latest one will be graded.
- LFD refers to the textbook "Learning from Data".
- Please read and follow submission instructions. No exception will be made to accommodate incorrectly submitted files.

---

1. [20 pts] Exercise 8.15 (e-Chap:8-38) in LFD.

2. [25 pts] Exercise 8.16 (e-Chap:8-42) in LFD. Please ignore part (c) and only do (a), (b), (d).

3. [25 pts] This question is related to the loss functions we discussed in class.

   a. Describe what are hinge loss, logistic regression loss, and 0-1 loss mathematically. Describe their similarities and differences using the unified picture we developed in class.

   b. By relying on the result in the above question, consider a point that is correctly classified and distant from the decision boundary. Why would SVM's decision boundary be unaffected by this point, but the one learned by logistic regression be affected?

4. [30 pts] **Support Vector Machine for Handwritten Digits Recognition:** You need to use the software package LIBSVM http://www.csie.ntu.edu.tw/~cjlin/libsvm/ to finish this assignment. Two functions ***svm_train()*** and ***svm_predict()*** from LIBSVM library will be used in this question. The package has already been included in the code folder. You only need to run 'make' command at the package location to use them. Please read the "LIBSVM tutorial" section in "Readme.txt" file carefully to understand how to use these functions. The handwritten digits files are in the "data" folder: train.txt and test.txt. The starting code is in the "code" folder. In the data file, each row is a data example. The first entry is the digit label ("1" or "5"), and the next 256 are grayscale values between -1 and 1. The 256 pixels correspond to a 16×16 image. You are expected to implement your solution based on the given codes. The only file you need to modify is the "solution.py" file. You can test your solution by running "main.py" file. Note that code is provided to compute a two-dimensional feature (symmetry and average intensity) from each digit image; that is, each digit image is represented by a two-dimensional vector. These features along with the corresponding labels should serve as inputs to your solution functions.

   (a) (10 points) Complete the ***svm_with_diff_c()*** function. In this function, you are asked to try different values of cost parameter c.

(b) (10 points) Complete the ***svm_with_diff_kernel()*** function. In this function, you are asked to try different kernels (linear, polynomial and radial basis function kernels).

(c) (10 points) Summarize your observations from (a) and (b) into a short report. In your report, please report the accuracy result and total support vector number of each model. A briefly analysis based on the results is also needed.

Deliverable: You should submit (1) a hard-copy report (along with your write-up for other questions) that summarizes your results and (2) the "solution.py" file to Blackboard.