CSC 240/440 - 2017 Spring: Homework 3

Hand in the hardcopy before the class on Mar. 21

Requirement

Due to the request from some students, the homework is posted online right now, but would be updated probably every week until it is formally released. (A) or (G) indicates questions for all or just graduate students. Undergraduate students are not required to do (G) questions, but they can get bonus points from that. **Please hand in the hardcopy of your homework before the class.**

The homework must be completed individually. However, you are encouraged to discuss the general algorithms and ideas with classmates in order to help you answer the questions. If you work with one or more other people on the general discussion of the assignment questions, please record their names over every question they participated.

However, the following behaviors will receive heavy penalties (lose all points and apply the honest policy explained in syllabus)

- explicitly tell somebody else the answers;
- explicitly copy answers or code fragments from anyone or anywhere;
- allow your answers to be copied;
- get code from Web.

Please also indicate how many late days you want to apply to your submission (Check the late policy in the Syllabus). All late submission without indicating late days or running out the late days cannot be accepted. For medical reasons, if the homework in time cannot be submitted on time, you have to submit the certificate with your homework.

1 (A) 2 points

Question 6 in Chapter 5.

2 (A) 2 points

Question 7 in Chapter 5.

3 (A) 2 points

Question 8 in Chapter 5.

4 (A) 2 points

Question 9 in Chapter 5.

5 (A) 2 points

Suppose the frequency of a disease in the population is 0.1%. There is a highly accurate screening test for this disease, which has a 5% false positive rate and a 10% false negative rate. (*False positive*: a result that indicates a given condition has been fulfilled, when it has not. *False negative*: a result that indicates a given condition has failed, while it was successful.) You take the test and the result is positive. What is the probability that you have the disease?

6 (A) 2 points

Given four training samples: $A = (1, 0), B = (-1, 0), C = (0, -1), D = (0, 1)$. A and B are in the positive class, while C and D are in the negative class. Assume to use the Euclidean distance to define the distance between any two points. We want to use K-NN to classify the training samples.

1. Let $K = 1$. Draw the decision boundaries based these four training samples and indicate the class label for each subarea.

2. Let $K = 3$. Draw the decision boundaries based these four training samples and indicate the class label for each subarea.

7 (A) KNN (4 points)

This is a programming question. You are asked to implement the K-NN. In this question, we are playing with an old data set but an interesting problem.

The Disputed Federalist Papers: The Federalist Papers were written in 1787-1788 by Alexander Hamilton, John Jay, and James Madison to persuade the citizens of the State of New York to ratify the U.S. Constitution. As was common in those days, these 77 essays, about 900 to 3500 words in length, appeared in newspapers signed with a pseudonym, in this instance, "Publius". In 1778, these papers together with eight additional articles on the same subject a total of 85 articles were published in book form. Since then, the consensus has been that John Jay was the sole author of 5 of a total 85 papers, that Hamilton was the sole author of 51, that Madison was the sole author of 14, and that Madison and Hamilton collaborated on another three. The authorship of the remaining 12 papers, known as the "disputed papers", has been a matter of long-standing controversy. It has been generally agreed that the disputed papers were written by either Madison or Hamilton, but there was no consensus about which were written by Hamilton and which by Madison. In this project, we look at the frequencies with which Madison and Hamilton used certain words, and use

this information to try to determine which of them wrote each of the 12 disputed papers. More background story can be found from http://en.wikipedia.org/wiki/The_Federalist_Papers.

In this homework, we look at the frequencies with which Madison and Hamilton used certain words, and use this information to try to determine which of them wrote each of the 12 disputed papers.

The data is available in the files "train_86_by_71.txt", "tune_20_by_71.txt", and "test_12_by_70.txt". The total number of samples is 118 (one line per paper). (A number of other papers with known authorship of either Hamilton or Madison were added to the dataset, to provide extra data on how the two authors made use of vocabulary.) The "train_86_by_71.txt" file is used to train your model (K-NN). The "tune_20_by_71.txt" is used to tune the parameter K. The first entry in each line contains the code number of the author: 1 for Hamilton and 2 for Madison. The remaining entries contain 70 floating point numbers that correspond to the relative frequencies (number of occurrences per 1000 words of the text) of the 70 function words, which are also available in the data file as an array of strings. The "train_86_by_71.txt" has 86 samples and the "tune_20_by_71.txt" file has 20 samples.

The "test_12_by_70.txt" file is used to test your K-NN model. It has 12 samples, that is, 12 disputed papers. Each line in this file corresponds to a disputed paper. Of course you have no labels on them.

You are required to implement the K-NN algorithm to classify the samples in The "test_12_by_70.txt" file. Your model can only be constructed from the training data set, that is, "train_86_by_71.txt". The tuning data set, that is, "tune_20_by_71.txt", can only be used for deciding the parameters in K-NN. The best value for K is decided by the best prediction accuracy on the tuning data set. Use your optimal value for K and the model learned from the training data set to predict the authorship for 12 disputed papers in file "test_12_by_70.txt".

Please report the following results:

- your best value K;
- the prediction accuracy on the tuning data set "tune_20_by_71.txt";
- the prediction result on the testing data set "test_12_by_70.txt".

You should upload your code to Blackboard and report your results as hard copy (hand in it together with your answers for other questions).

Important Reminder: To obtain a reasonable accuracy, you should consider to normalize the all attributes. Consider an attribute for n samples (including all samples your have including training, validation, and testing), say, $a_1, a_2, \cdots, a_n$. The new values would be

$$a'_i = \frac{a_i - \mu}{\sigma} \quad \text{for all} \quad i = 1, \cdots, n$$

where

$$\mu = \frac{1}{n}\sum_{i=1}^{n} a_i \quad \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - \mu)^2}$$

8 (A) SVM (4 points)

This is a programming question. You are asked to use the SVM classifier to predict the authorship in the last problem. Recall that the SVM formulation is

$$\min_{w,b} \quad \sum_{n=1}^{N} \max(0, 1 - y_i(\mathbf{x}_i^\top \mathbf{w} + b)) + \frac{C}{2}\|\mathbf{w}\|^2$$

We recommend you to use libraries rather than implementing everything from scratch (e.g., the Python based Scikit[1], LIBSVM[2] or LIBLINEAR[3]).

Tuning the parameter C correctly is a vital step in the use of SVM. The C parameter trades off misclassification of training examples against simplicity of the decision surface. A low C makes the decision surface smooth, while a high C aims at classifying all training examples correctly by giving the model freedom to select more samples as support vectors.

Please report the following results:

- your best value C;
- the prediction accuracy on the training data set "train_86_by_71.txt";
- the prediction accuracy on the tuning data set "tune_20_by_71.txt";
- the prediction result on the testing data set "test_12_by_70.txt".

You should upload your code to Blackboard and report your results as hard copy (hand in it together with your answers for other questions).

9 (A) Linear regression (4 points)

This is a programming question. You are asked to use the linear regression (LR) to predict the authorship in the last problem. Just treat the label as continuous value.

Please report the following results:

- How do you predict the label by using LR model? Write down your idea.
- the prediction accuracy on the training data set "train_86_by_71.txt";
- the prediction accuracy on the tuning data set "tune_20_by_71.txt";
- the prediction result on the testing data set "test_12_by_70.txt".

You should upload your code to Blackboard and report your results as hard copy (hand in it together with your answers for other questions).

[1] `http://scikit-learn.org/stable/modules/svm.html`
[2] `https://www.csie.ntu.edu.tw/~cjlin/libsvm/`
[3] `https://www.csie.ntu.edu.tw/~cjlin/liblinear/`

10 (A) ANN (6 points)

This is a programming question. You are asked to train an artificial neural network (ANN) for the task of handwritten digit recognition.

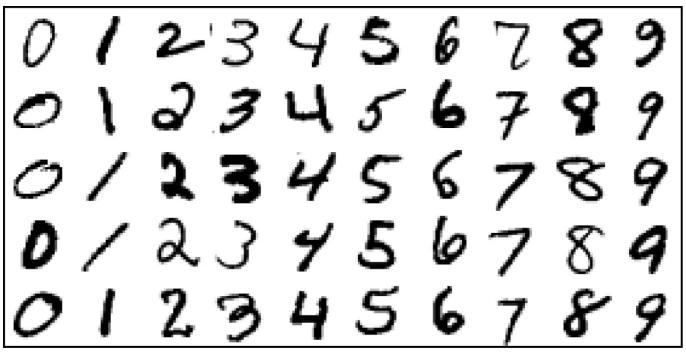

Figure 1: Handwritten Digits

The data files "train_digit.csv", "tune_digit.csv" and "test_digit.csv" contain gray-scale images of hand-drawn digits, from zero through nine. Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive. Each data set contains 785 columns. The first column (Column 0) is the label: the digit drawn by the user. The rest of the columns contain the pixel-values of the associated image in the following way:

$$\begin{bmatrix} \text{Col.1} & \text{Col.2} & \text{Col.3} & \dots & \text{Col.28} \\ \text{Col.29} & \text{Col.30} & \text{Col.31} & \dots & \text{Col.56} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Col.757} & \text{Col.758} & \text{Col.759} & \dots & \text{Col.784} \end{bmatrix}$$

We recommend you to use existing solvers, for example, Python based Scikit[4], the Neural Network Toolbox in Matlab, or R. Decide the optimal network structure based on the accuracy on the validation data set.

Please report the following results:

- your ANN structure;
- the prediction accuracy on the training data set "train_digit.csv";
- the prediction accuracy on the tuning data set "tune_digit.csv";
- the prediction result on the testing data set "test_digit.csv".

You should upload your code to Blackboard and report your results as hard copy (hand in it together with your answers for other questions).

[4] `http://scikit-learn.org/stable/modules/neural_networks_supervised.html`