Homework 3
CS 436/580L: Introduction to Machine Learning

Instructor: Arti Ramesh

Instructions

1. You can use either C/C++, Java or Python to implement your algorithms.
2. **Your implementations should compile on remote.cs.binghamton.edu.**
3. Make sure remote.cs.binghamton.edu has the packages that you require before starting to implement.
4. This homework requires you to **implement** Logistic Regression and Perceptrons. Using existing packages for both these algorithms is not allowed.
5. Please make sure your code is readable and well-commented.
6. Your homework should contain the following components:
   (a) README.txt file with detailed instructions on how to compile and run the code.
   (b) Code source files
   (c) Type-written document containing the results on the datasets.
7. Submit the homework as a **single zip file:** $firstname\_lastname\_hw3.zip$.

1 Logistic Regression

0 points Download the datasets available on myCourses. As in homework 2, the classification task is spam/ham.

25 points Implement the MCAP Logistic Regression algorithm with L2 regularization that we discussed in class (see Mitchell's new book chapter). Try five different values of λ (constant that determines the strength of the regularization term). Use your algorithm to learn from the training set and report accuracy on the test set for different values of λ. Implement gradient ascent for learning the weights. Do not run gradient ascent until convergence; you should put a suitable hard limit on the number of iterations.

5 points Improve your Logistic Regression algorithms by throwing away (i.e., filtering out) stop words such as "the" "of" and "for" from all the documents. A list of stop words can be found here: `http://www.ranks.nl/resources/stopwords.html`. Report accuracy for Logistic Regression for this filtered set. Does the accuracy improve? Explain why the accuracy improves or why it does not?

2 Perceptrons, and Neural Networks

In this question, you will implement the Perceptron algorithm and compare it with WEKA implementation of Neural networks. You will also compare it with your own implementations of Logistic Regression and Naive Bayes. If you are unsure of whether your implementation is correct, then you may compare them with WEKA implementations of Logistic Regression and Naive Bayes.

35 points Implement the perceptron algorithm (use the perceptron training rule and not the gradient descent rule). Your task here is to experiment with different values of number of iterations and the learning rate. Report the accuracy for 20 suitable combinations of number of iterations and the learning rate. Repeat your experiment by filtering out the stop words. Compare the accuracy of your perceptron implementation with that of Naive Bayes (implemented in Homework 2) and Logistic Regression (implemented in this homework).

10 points Neural networks in WEKA.

- Download WEKA `http://www.cs.waikato.ac.nz/ml/weka/`.
- Convert the spam/ham dataset into the ARFF format used by WEKA.
- Using the Neural networks implementation in WEKA (called MultiLayered Perceptron), report the accuracy on the test set. Experiment with different number of hidden layers and units. Report on how the number of hidden layers and units as well as other options such as momentum, number of iterations, and learning rate affect the accuracy.

What to Turn in

- Your code
- **(5 points)** README file for compiling and executing your code.
- **(10 points)** A detailed write up that contains:
    1. The accuracy obtained on the test set using Logistic Regression for different values of λ.
    2. The accuracy on the test set after filtering the stop words.
    3. The accuracy on the test set different values of the number of iterations and the learning rate.
    4. The accuracy on the test set different number of hidden layers and units, momemtum, number of iterations, and learning rate.
    5. Compare the accuracy across the different models and report your observations.