Machine Learning: Homework 3

Due 11:55 p.m. Friday, March 14, 2020

Instructions

- **Collaboration policy:** Homeworks must be done individually, except where otherwise noted in the assignments. "Individually" means each student must hand in their own answers, and each student must write and use their own code in the programming parts of the assignment. It is acceptable for students to collaborate in figuring out answers and to help each other solve the problems, though you must in the end write up your own solutions individually, and you must list the names of students you discussed this with. We will be assuming that, as participants in an undergraduate course, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.
- **Format of Submission:** Please submit your homework with a single zip file. Files in the zip file are
  - **NetID-HW3.pdf** : a write-up file which contains answers to problems, and **it should contain the commands for running your code**;
  - **problem-X.py** : the code files for problems (where X is the ID of the problem).
- **Online submission:** You must submit your solutions online on NYU Classes. The write-up file should be in PDF format. We recommend that you use LaTeX. PDF files exported from doc/docx files are also accepted. Please do not submit hand-written solutions.
- **skeleton code:** We provide skeleton code files for Problem 1 and Problem 3. Please use them.

Problem 1: Naïve Bayes

Using the same spam email dataset as we used in Homework 1, implement a Naïve Bayes classifier for spam email classification. See **naive-bayes-skeleton.py** for the skeleton code.

(a) **[5 Points]** Use boolean features as we did in Homework 1, i.e., x_j being 1 if the j^{th} word in the vocabulary occurs in the email, or 0 otherwise.

(b) **[5 Points]** Plot the training and validation errors as a function of training size N. (Hint: This is the same as what you did in Homework 1, Problem 1.5. Also, you may use log-scale to avoid numerical issues.)

(c) **[5 Points]** To improve the accuracy, feel free to change the dimensionality of features by using a different vocabulary threshold X. Tell me about what you try and what result you get.

(d) **[5 Points]** Try both MLE and MAP estimators (For simplicity, just set the hallucinated word count to be 1). Which one is better? Please briefly explain. (For the previous two questions, you only have to report the result of either MLE or MAP, whichever is better according to your experiment.)

Problem 2 (Bonus): Naïve Bayes with Bag of Words

[+10 Points] Repeat Problem 1 using the "bag of words" features with Naïve Bayes classifier.

Problem 3: Gaussian Naïve Bayes with Bag of Words

[**20 Points**] Repeat Problem 1 using the "bag of words" features with **Gaussian** Naïve Bayes classifier. See **gaussian-nb-skeleton.py** for the skeleton code.

Problem 4: Comparison

[**5 Points**] Now you have mastered three algorithms to do spam email classification: Perceptron, Naïve Bayes, and Gaussian Naïve Bayes. Briefly summarize their pros and cons based on your understanding and experiment. (Keep it less than 150 words.)