CSCC11 – Introduction to Machine Learning and Data Mining

Fall 2022

Assignment 2

Logistics

- This assignment is due on Nov 17, 2022 at 11:59 pm.
- It can be done individually or in groups of two.
- Some questions require written answers while others require writing Python code. Make sure to upload the various components of your solution as a single archive file (e.g. rar, zip) to "Assignment 2" on Quercus. The file should be named using the student IDs of the group members.
- The answers/code you submit must be your own.
- We prefer that you ask any related questions during the tutorial sessions or through Piazza. Otherwise, for your first point of contact, please reach out to your TAs:

  Dhiraj Tawani, through email: dhiraj.tawani@mail.utoronto.ca
  Srinath Dama, through email: srinath.dama@mail.utoronto.ca

  Any such inquiries should be done by Nov 16th at the latest.

Part 1: Written Component [40 points]

Assume we want to build a logistic regression model to classify a fruit as orange/non-orange using its width and height. The training data to be used is as follows:

Width	Height	Orange
4	4	Yes
6	4	Yes
6	5	Yes
6	8	No
6	10	No
8	8	Yes
8	10	No

a) Write the corresponding optimization problem in terms of the data provided above and specify the parameters to be estimated.

b) Perform 3 iterations of the steepest descent algorithm to determine the parameters assuming that the initial estimate is $[0.3, -0.2, 0.7]^T$ and the step size (λ) is 0.01. For each estimate (including the initial one), you are required to report the following:

   - The value of the estimate
   - The accuracy of the resulting logistic regression model when applied to the training data

   Note that you do not need to do the computations manually. You might want to use a spreadsheet or write a simple code to do that.

c) Classify the following data points using the model you obtained in part b: (3, 3), (4, 10), (9, 8), and (9, 10).

d) Discuss one advantage of logistic regression.

e) Briefly explain whether logistic regression is discriminative or generative.

Part 2: Programming Component* [60 points]

Using this dataset from the BBC, you are required to implement and evaluate different classifiers to label news articles according to five categories: business, entertainment, politics, sport, and tech. Your code should not make use of any existing implementation of these classifiers.

The first step is to download the pre-processed dataset and inspect the different files to understand how to read the information in your code:

- The dataset includes 2225 articles listed in the *bbc.docs* file. The actual articles can be found here.

- The data has been pre-processed using stemming, stop-word removal, and low term frequency filtering. This resulted in 9635 terms listed in the *bbc.terms* file.

- Each row in *bbc.mtx*, except the first two, represents the frequency of a term in a given article. For example, row 812 (“2 528 5.0”) indicates that term 2 (“sale”) occurs 5 times in article 528 (*entertainment.018*).

* The original version of this question was prepared by Prof. Francisco Estrada

a) Write a Python code that does the following:
    i. Reads the dataset and partitions it into training and testing subsets.
    ii. Trains a Naïve Bayes classifier using the training subset. Note that you need to represent the occurrence of each term as a binary value instead of a frequency. In addition, to avoid dealing with zero probabilities for words that do not occur in a particular class of articles, you can adjust the conditional probabilities $p(x_i|y)$ by adding 1 to the numerator and 2 to the denominator.
    iii. Determines the classification accuracy by computing the percentage of labels returned by your classifier that agrees with the labels attached to the dataset. This step should be done for the training and testing subsets separately. In other words, you need to compute the accuracy of your classifier when applied to the training subset (i.e. when used to predict the labels of the training data points) and then do the same for the testing subset. Keep in mind that the classifier should be built using the training subset only.

b) Repeat the steps of part a) assuming a Gaussian class conditionals classifier with frequency-based features instead of binary ones. For this part, you need to consider five Gaussian components (one for each category of articles) whose parameters should be determined using Maximum Likelihood Estimation.

c) Repeat the steps of part a) using k-Nearest Neighbors classification. You are required to consider binary as well as frequency-based features and report classification accuracy for the following values of k: 1, 3, and 6. That is, there should be a total of six accuracy results (one for each combination of feature type and k value). Note that, in the case of ties, you can randomly select one of the tied labels.