CS 4375 Intro to Machine Learning Dr. Karen Mazidi

Portfolio Component: ML Algorithms from Scratch

Pair or Individual Assignment

Objectives:

- Gain a deeper understanding of machine learning algorithms by coding from scratch
- Further C++ skills in ML
- Understand the importance of reproducible research

Turn in:

- Your two C++ programs and the report to eLearning and to your portfolios, creating links on your portfolio
- Only one person in a team needs to upload to eLearning, but both need to post everything on your GH repos.

Instructions:

1. Program 1. In the C++ IDE of your choice:
    a. Read in the Titanic data, as you did in Portfolio Assignment Data Exploration.
    b. Write a program to perform logistic regression from scratch as discussed in class. Predict survived based on sex, ignoring the other predictors. Use the first 800 observations for train. You can use the R code in Chapter 6 as pseudocode. Output your coefficients. To check if the coefficients are correct, you can run logistic regression in R on the same training data.
    c. Use the remaining data to predict values.
    d. Write functions to calculate accuracy, sensitivity, specificity.
    e. Output the test metrics and the run time for the algorithm. You can use chrono to measure time. Measure just the training time of the algorithm.
2. Program 2. Repeat the above but implement Naïve Bayes using predictors age, pclass, and sex to predict survival on the Titanic data. Use the first 800 observations for train, the rest for test.
3. Write a document:
    a. copy/paste runs of your code showing the output (coefficients and metrics), and run times
    b. analyze the results of your algorithms on the Titanic data
    c. write two paragraphs comparing and contrasting generative classifiers versus discriminative classifiers. Cite any sources you use.
    d. Google this phrase: reproducible research in machine learning. Using 2-3 sources, at least one of which should be academic, write a couple of paragraphs of what this means, why it is important, and how reproducibility can be implemented. Cite your sources using any format.
4. Create a link to this document and your code on your index page.

Caution: All course work is run through plagiarism detection software comparing students' work as well as work from previous semesters and other sources.

Notes:

- be prepared to demo your code to the class
- feel free to use whatever data structures you like: arrays, vectors, etc.
- for matrix multiplication, it's faster if you code it by hand but feel free to use a library like armadillo or eigen (each has its own learning curve)
- Here is a great video that gives a conceptual picture of naïve Bayes with Gaussian predictors: https://www.youtube.com/watch?v=r1in0YNetG8
- The following formula shows how to calculate the likelihood of a continuous predictor. The book gives hints as well..

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

Grading Rubric:

Element	Points
Step 1 C++ code	150
Step 2 Overview document	40
Step 3 Create links to the document and code on the index page	10
Total	200

- 90% and above for exceptional work
- 80-89% for good work
- 70-79% for average work
- below 70% for low quality work

Caution: All course work is run through plagiarism detection software comparing students' work as well as work from previous semesters and other sources.