Homework 2: Classification With Support Vector Machines

About

Due

~~Monday 2/4/19, 11:59 PM CST~~

Wednesday 2/6/19 11:59 PM CST (Updated 2/2)

Goal

This homework focuses on implementing and using a support vector machine for classification.

Additionally, we look at using a validation set to tune parameters, specifically, the learning rate of the SVM.

Code and External Libraries

The assignment can be done using any language.

External libraries should NOT be used to train the SVM, perform train-test splits, or parameter tuning. However, external libraries can be used to load and manipulate the data.

Problems

Total points: 100

The UC Irvine machine learning data repository hosts a collection of data on adult income, donated by Ronny Kohavi and Barry Becker. You can find this data at https://archive.ics.uci.edu/ml/datasets/Adult (https://archive.ics.uci.edu/ml/datasets/Adult) For each record, there is a set of continuous attributes, and a class "less than 50K" or "greater than 50K". We have pre-split the data training (./train.txt) with 43957 examples with known class labels, and testing (./test.txt) data with 4885 examples without class labels. Use this data, not the original, for this assignment.

Write a program to train a support vector machine on this data using stochastic gradient descent, as detailed in Procedure 4.3 from the text.

(Update 2/2) As several people pointed out, the gradient formula provided during AML lecture was incorrect. The correct formulas can be found in the text under the heading "Setting up stochastic gradient descent". Specifically, the correct gradients are:

$$\nabla_a = \begin{cases} \lambda a & if\ y_k(a^T x_k + b) \geq 1 \\ \lambda a - y_k x_k & otherwise \end{cases}$$

$$\nabla_b = \begin{cases} 0 & if\ y_k(a^T x_k + b) \geq 1 \\ -y_k & otherwise \end{cases}$$

You should not use a package to train the classifier (that's the point), but your own code. You should ~~ignore the id number, and~~ use **only** the continuous variables as a feature vector. You should scale these variables so that each has unit variance, and you should subtract the mean so that each has zero mean. You should search for an appropriate value of the regularization constant, trying at least the values [1e-3, 1e-2, 1e-1, 1]. Use 10% of your training data as a validation set for this search. You should use at least 50 ~~epochs~~ seasons of at least 300 steps each. In each ~~epoch~~ season, you should separate out 50 training examples at random for evaluation (call this the set held out for the ~~epoch~~ season). You should compute the accuracy of the current classifier on the validation set for the ~~epoch~~ season every 30 steps.

You should produce:

- A plot of the validation accuracy every 30 steps, for each value of the regularization constant.
- A plot of the magnitude of the coefficient vector every 30 steps, for each value of the regularization constant.
- Your estimate of the best value of the regularization constant, together with a brief description of why you believe that is a good value.
- Answer the question: What was your choice for the learning rate and why did you choose it?
- Once you have trained your final classifier, score the provided test set, recording the results in a file with the same format as submission.txt (./submission.txt). You will be able to submit this result to gradescope repeatedly for scoring.

Submission

Your submission should be a PDF with the following pages.

Page 1 A screenshot of your best accuracy on the test set (retrieve this from the autograder).

Page 2 A plot of the validation accuracy every 30 steps, for each value of the regularization constant. You should plot the curves for all regularization constants in the same plot using different colors with a **label showing the corresponding values**.

Page 3 A plot of the magnitude of the coefficient vector every 30 steps, for each value of the regularization constant. You should plot the curves for all regularization constants in the same plot using different colors with a **label showing the corresponding values**.

Page 4 Your estimate of the best value of the regularization constant, together with a brief description of why you believe that is a good value. What was your choice for the learning rate and why did you choose it?

Page 5 A screenshot of your code.

The page should contain snippets of code demonstrating:

- Training of an SVM, including but not limited to SGD.
- Testing of an SVM.

Page 6+

All code should be attached at the end of the pdf. There is no limit to the number of pages required for full code printout.