CS 613 - Machine Learning

Assignment 2 - Logistic Regression

Introduction

In this assignment you will implement a binary Logistic Regression classifiers for the purpose of binary and multi-class classification.

You may **not** use any functions from an ML library in your code. And as always your code should work on any dataset that has the same general form as the provided one.

Grading

Although all assignments will be weighed equally in computing your homework grade, below is the grading rubric we will use for this assignment:

Part 1 (Theory)	10pts
Part 2 (Binary Logistic Regression)	50pts
Part 3 (Multi-Class Logistic Regression)	40pts
TOTAL	100 pts

Datasets

Spambase Dataset (spambase.data) This dataset consists of 4601 instances of data, each with 57 features and a class label designating if the sample is spam or not. The features are *real valued* and are described in much detail here:

https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.names

Data obtained from: https://archive.ics.uci.edu/ml/datasets/Spambase

Iris Dataset (iris.data) This dataset consists of 150 instances of data, each with four features and a class label. The features are *real valued* and are described in much detail here:

https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.names

Data obtained from: https://archive.ics.uci.edu/ml/datasets/iris

1 Theory

1. For the function $J = (x_1 w_1 - 5 x_2 w_2 - 2)^2$, where $w = [w_1, w_2]$ are our weights to learn:

   (a) What are the partial gradients, $\frac{\partial J}{\partial w_1}$ and $\frac{\partial J}{\partial w_2}$? Show work to support your answer (6pts).

   (b) That are the value of the partial gradients given current values of $w = [0, 0], x = [1, 1]$ (4pts)?

2 Spambase Logistic Regression Classier

For your first programming task, you'll implement, train and validate a *Logistic Regression Classifier* for determining if an email is spam (or not).

First download the dataset *spambase.data* from Blackboard. As mentioned in the Datasets area, this dataset contains 4601 rows of data, each with 57 continuous valued features followed by a binary class label (0=not-spam, 1=spam). As always, your code should work on any dataset that lacks header information and has several comma-separated continuous-valued features followed by a class id $\in 0, 1$.

Write a script that:

1. Reads in the data.
2. Shuffles the observations.
3. Selects the first 2/3 (round up) of the data for training and the remaining for validation.
4. Zscores the features based on the training data.
5. Trains a logistic regression model using gradient descent, keeping track of the mean log loss of the training and validation sets as it trains.
6. Plots the training and validation mean log loss as a function of the epoch.
7. Computes the precision, recall, f-measure and accuracy of the learned model on the training and validation sets when using a threshold of 0.5.
8. Plots a precision-recall graph by varying the threshold from 0.0 to 1.0, inclusive, in increments of 0.1.

Implementation Details

1. Seed the random number generate with zero prior to randomizing the data.
2. You may choose how to initialize your parameters (weights and bias).
3. You may choose your learning rate.
4. You may choose your termination criteria(s).

In your report you will need:

1. The plot of training and validation log loss as a function of the epoch.
2. The requested statistics on the training and validation sets.
3. The requested precision-recall graph.

3 Logistic Regression for Multi-Class Classification

We'll now use multiple logistic regression models to perform multi-class classification to classify an Iris flower to be one of three types.

First download the dataset *iris.data* from Blackboard. As mentioned in the Datasets area, this dataset contains 150 rows of data, each with four continuous valued features followed by a categorical nominal class label.

Write a script that:

1. Reads in the data.
2. Shuffles the observations.
3. Selects the first 2/3 (round up) of the data for training and the remaining for validation.
4. Zscores the features based on the training data.
5. Trains three models for one-vs-one multi-class classification, each of which:
   (a) Selects the samples pertaining to the two classes your comparing.
   (b) Trains a logistic regression model using gradient descent.
6. Applies the models to each validation sample to determine the most likely class.
7. Computes the validation accuracy.
8. Creates a confusion matrix for the validation data.

Implementation Details

1. Seed the random number generate with zero prior to randomizing the data.
2. You may choose how to initialize your parameters (weights and bias).
3. You may choose your learning rate.
4. You may choose your termination criteria(s).
5. For deciding which class to assign on observation, compute the mean likelihood for each class (using the likelihoods provided by the models they're in), and **choose the class with the largest mean likelihood**.
6. You must compute the values for the confusion matrix yourself (not using some library function). The cells of the confusion matrix can just have the total counts pertaining to their location.

In your report you will need:

1. The validation accuracy.
2. Your confusion matrix for the validation data.

Submission

For your submission, upload to Blackboard a single zip file (again no spaces or non-underscore special characters in file or directory names) containing:

1. PDF Writeup
2. Source Code
3. readme.txt file

The readme.txt file should contain information on how to run your code to reproduce results for each part of the assignment.

The PDF document should contain the following:

1. Part 1: Answers to theory question(s).
2. Part 2:
   (a) The plot of training and validation log loss as a function of the epoch.
   (b) The requested statistics on the training and validation sets.
   (c) The requested precision-recall graph.
3. Part 3:
   (a) The validation accuracy.
   (b) Your confusion matrix for the validation data.