CS 383 - Machine Learning

Assignment 4 - Classification

Introduction

In this assignment you will implement a Naive Bayes classifier for the purpose of binary 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)	55pts
Part 2 (Naive Bayes)	35pts
Report	10pts
Extra Credit	10pts
TOTAL	110 (of 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

1 Theory

1. Consider the following set of training examples for an unknown target function: $(x_1, x_2) \rightarrow y$:

Y	x_1	x_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

   (a) What is the sample entropy, $H(Y)$ from this training data (using log base 2) (5pts)?

   (b) What are the information gains for branching on variables x_1 and x_2 (5pts)?

   (c) Draw the decision tree that would be learned by the ID3 algorithm without pruning from this training data. All leaf nodes should have a single class choice at them. If necessary use the mean class or, in the case of a tie, choose one at random.(10pts)?

2. We decided that maybe we can use the number of characters and the average word length an essay to determine if the student should get an A in a class or not. Below are five samples of this data:

# of Chars	Average Word Length	Give an A
216	5.68	Yes
69	4.78	Yes
302	2.31	No
60	3.16	Yes
393	4.2	No

   (a) What are the class priors, $P(A = Yes), P(A = No)$? (5pts)

   (b) Find the parameters of the Gaussians necessary to do Gaussian Naive Bayes classification on this decision to give an A or not. Standardize the features first over all the data together so that there is no unfair bias towards the features of different scales (5pts).

   (c) Using your response from the prior question, determine if an essay with 242 characters and an average word length of 4.56 should get an A or not. Show the math to support your decision (10pts).

3. Another common activation function for use in logistic regression or artificial neural networks is the hyperbolic tangent function, *tanh*, which is defined as:

$$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{1}$$

(a) Since the hyperbolic tangent function outputs values in the range of $-1 <= tanh(z) <= 1$ we will have to augment our log likelihood objective function to deal with this range. If we opt to use this function for logistic regression (as opposed to $\frac{1}{1+e^{-}z}$), what will this object function be? Show your work. (5pts)

(b) In order to compute the gradient of your previous answer with respect to θ_j, we'll need to compute the gradient of the hyperbolic tangent function itself. Use the exponential definition of the hyperbolic tangent function provided at the top of this problem to show that
$\frac{\partial}{\partial\theta_j}(tanh(x\theta)) = x_j(1 - tanh(x\theta)^2)$. (5pts)

(c) Using the fact that $\frac{\partial}{\partial\theta_j}(tanh(x\theta)) = x_j(1 - tanh(x\theta)^2)$, what is the gradient of your log likelihood function in part (a) with respect to θ_j? Show your work. (5pts)

2 Naive Bayes Classifier

For your first programming task, you'll train and test a *Naive Bayes Classifier*.

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). There is no header information in this file and the data is comma separated. 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. Randomizes the data.
3. Selects the first 2/3 (round up) of the data for training and the remaining for testing
4. Standardizes the data (except for the last column of course) using the training data
5. Divides the training data into two groups: Spam samples, Non-Spam samples.
6. Creates Normal models for each feature for each class.
7. Classify each testing sample using these models and choosing the class label based on which class probability is higher.
8. Computes the following statistics using the testing data results:
   (a) Precision
   (b) Recall
   (c) F-measure
   (d) Accuracy

Implementation Details

1. Seed the random number generate with zero prior to randomizing the data
2. Matlab interprets $0log0$ as NaN (not a number). You should identify this situation and consider it to be a value of zero.

In your report you will need:

1. The statistics requested for your Naive Bayes classifier run.

Precision:	Around 68%
Recall:	Around 95%
F-Measure:	Around 79%
Accuracy:	Around 81%

Table 1: Evaluation for Naive Bayes classifier

3 Extra Credit: The Precision-Recall Tradeoff

For 10 extra credit points, find a data set of your choosing on which you can perform binary classification. Now apply your Naive Bayes code to this dataset and vary the threshold required to make an observation be considered your positive class.

Write a script that:

1. Reads in the data.
2. Randomizes the data.
3. Selects the first 2/3 (round up) of the data for training and the remaining for testing
4. Standardizes the data using the training data
5. Divides the training data into two groups: Positive samples, Negative samples.
6. Creates Normal models for each feature and each class.
7. Computes the $P(Positive|data)$ and $P(Negative|data)$ for each testing sample using Naive Bayes, normalizing them such that $P(Positive|data) + P(Negative|data) = 1$.
8. **Vary the threshold from 0.0 to 1.0 in increments of 0.05**, each time:
   (a) Using the current threshold, label each testing sample as True Positive, True, Negative, False Positive, or False Negative
   (b) Compute the Precision and Recall for this threshold level.
9. Plot Precision vs Recall.

Implementation Details

1. In computing Precision, Recall, F-Measure and Accuracy make sure the denominators don't become zero. If they do, check the numerator. If that's also zero then set the value to one.

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. If you did the extra credit, the dataset used.
4. 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:
   (a) Answers to theory questions
2. Part 2:
   (a) Requested Classification Statistics
3. Extra Credit:
   (a) Citation and link to dataset.
   (b) Plot of Precision-vs-Accuracy