CS 383 - Machine Learning

Assignment 4 - Naive Bayes, Decision Trees and Nearest Neighbors
Summer 2017

Introduction

In this assignment you will perform classification using the Nearest Neighbors algorithms in addition to theoretical questions related to Naive Bayes and decision trees.

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)	22pts
Part 2 (K-Nearest Neighbors)	20pts
Report	5pts
TOTAL	47pts

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) (2pts)?

   (b) What are the information gains for branching on variables x_1 and x_2 (4pts)?

   (c) Draw the deicion tree that would be learned by the ID3 algorithm without pruning from this training data (5pts)?

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, $P(A = Yes), P(A = No)$? (1pt)

   (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 (5pts).

2 k-Nearest Neighbors (KNN)

Throughout this assignment we are going to train systems for the task of classifying instances as spam or not spam. The first system you'll implement will be *k-Nearest Neighbors*

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). 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. Performs k-Nearest Neighbors classification with $k = 5$.
6. Computes the following statistics using the testing data:
   (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. For your distance measurement use the Manhattan (L1) distance.
3. Set $k = 5$ but make this a variable so you could easily change it.
4. For choosing the class, use the majority class (the mode) of the neighbors.

In your report you will need:

1. The statistics requested for your kNN run.

Precision:	$\approx 92\%$
Recall:	$\approx 84\%$
F-Measure:	$\approx 88\%$
Accuracy:	$\approx 91\%$

Table 1: Evaluation for k-Nearest Neighbors with k=5

Submission

For your submission, upload to Blackboard a single zip file 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:
    (a) Answers to theory questions
2. Part 2:
    (a) Requested Classification Statistics