COMP 4106 - Artificial Intelligence
Winter 2019

Assignment #3

Due date: April 9, 2019

1 Bayesian, Decision Tree and Dependence Tree Classifiers

1.1 Introduction

In this assignment you will be implementing a few classification algorithms including the optimal Bayesian classifier, one for Decision Trees (DTs), and one for Dependence Trees, and using them to classify several different data sets.

1.2 Binary-valued *Artificial* Data Sets

1.2.1 Data Generation

Use the scheme below to generate the data sets you need:

1. You are dealing with a d-dimensional feature space with $c = 4$ classes. You can assume that $d = 10$.
2. Assume that the vector components obey a Dependence Tree structure between the various features. This Dependence Tree must be arbitrarily assigned and unknown to the classification (i.e., training and testing) algorithm.
3. For each of the c classes and for each of the d features, randomly generate the probabilities of the feature taking the value 0 or 1. Thus, for class $j = 1, \ldots, c$ and for feature indices $i = 1, \ldots, d$, you must randomly assign the value $v_{i,j} = Pr[x_i = 0|\omega = \omega_j]$. *These values must be based on the Dependence Tree that you have chosen.*
4. Generate 2,000 samples for each class based on the above features.

1.2.2 Training and Testing

With regard to training and testing, do the following:

1. Use a 5-fold cross-validation scheme for training and testing.
2. Using estimates of the $v_{i,j}$'s, estimate the true but unknown Dependence Tree. Record the results of how good your estimate of the true but unknown Dependence Tree is.
3. Perform a Bayesian classification[1] assuming that all the random variables are *independent*. Notice that in this case, you must not assume a Gaussian distribution for the features, but the *binary* distribution.
4. Perform a Bayesian classification assuming that all the random variables are *dependent* based on the dependence tree that you have inferred.
5. Perform the classification based on a DT algorithm. For the DT algorithm, have your program output the resulting DT. The output[2] should be neatly indented for easy viewing.

[1] Each data sets has more than two classes. In each case, you must do the classification using a pairwise classification on all the classes and assign the testing sample to the most appropriate "winning" class. This paradigm must be followed for the other classification tasks too.

[2] An excellent program to draw decision trees is Graphviz, available at: http://www.graphviz.org/.

1.3 Binary-valued *Real-life* Data Sets

In this section you will deal with the one *Real-life* data set.

1.3.1 Data

The Glass Identification data set[3] is to be used to classify the type of glass, given the following features, specified in this order:

1. Class: In this case there are 7 possible types, which can be further split in to 2 categories of windowed and non-windowed glass
2. Id: Number
3. RI: Refractive index
4. Na: Sodium (unit measurement is weight percent in the oxide, as are attributes 5-11)
5. Mg: Magnesium
6. Al: Aluminum
7. Si: Silicon
8. K: Potassium
9. Ca: Calcium
10. Ba: Barium
11. Fe: Iron

You may ignore all the features that are non-numeric. Whenever you need *binary* features (i.e., for training and classifying using the Dependence Tree and Decision Tree), render the features to be binary by adopting a thresholding mechanism.

1.3.2 Techniques to be Implemented

Perform all the tasks given in Section 1.2.2 on this real-life data set.

2 Report

1. Write a 2-3 page report summarizing all your results. The report should be relatively formal.
2. Compare the classification accuracy of the Dependence Trees you have obtained for the artificial and real-life data sets.
3. Compare the classification accuracy of the four algorithms for the artificial data sets. Do some seem to outperform others? Discuss the possible reasons for these results.
4. Compare the classification accuracy of the four algorithms ((a) Bayes, (b) Naive Bayes, (c) using Dependence trees, and (d) using Decision Trees) for the real-life data sets. Do some seem to outperform others? Again, discuss the possible reasons for these results.

[3]This data set can be found at the UCI Machine Learning Repository. It is located at https://archive.ics.uci.edu/ml/machine-learning-databases/glass/.