COMP 4106 - Artificial Intelligence
Winter 2015

Assignment #3

Due date: March 27, 2015

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 8-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.*

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 following three *Real-life* data sets available on the course website as Datasets.zip.

1.3.1 Data

The Iris data set will be used to classify the type of the Iris flower, given the following features in this order:

1. Sepal length in cm
2. Sepal width in cm
3. Petal length in cm
4. Petal width in cm
5. Class: Iris Setosa, Iris Versicolour, or Iris Virginica

The Wine data set will be used to classify the type of Wine given the following features in this order:

1. Class: In this case there are three classes.
2. Alcohol
3. Malic acid
4. Ash
5. Alcalinity of ash
6. Magnesium
7. Total phenols
8. Flavanoids
9. Nonflavanoid phenols
10. Proanthocyanins
11. Color intensity
12. Hue
13. OD280/OD315 of diluted wines
14. Proline

The Heart Disease data set will be used to classify the presence of heart disease given the following features in this order:

1. Age
2. Gender
3. Cp
4. Trestbps
5. Chol

6. Fbs
7. Restecg
8. Thalach
9. Exang
10. Oldpeak
11. Dlope
12. Ca
13. Thal
14. Classes: 1 - No presence; 2 - Cond. 1; 3 - Cond. 2; 4 - Cond. 3; 5 - Cond. 4

In all the above cases, ignore all the features that are non-numeric. 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 these real-life data sets.

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 for the real-life data sets. Do some seem to outperform others? Again, discuss the possible reasons for these results.