University of Agder

IKT 715 - Advanced Topics in Pattern Recognition
Winter 2019

Cumulative Assignment for the Course

The Tasks are On the Topics Taught Each Day

Bayesian, Anti-Bayesian, Decision Tree, and Dependence Tree Classifiers

1 Introduction

This assignment is more realistically a "Project" for the entire course. It can be considered as a "Cumulative Assignment" which will train you in all the advanced topics that you are taught.

In this assignment you will implement a few classification algorithms including the optimal Bayesian classifier, one for the Anti-Bayesian classifier, one for Decision Trees (DTs), and one for Dependence Trees (DepTs). You will then use them to classify several different data sets.

2 Training/Testing Methodology

For all the schemes requested below, use a 5-fold cross-validation scheme for training and testing. Also, each data set has more than two classes. So, you must do the classification using a pairwise decision on all the classes, and assign the testing sample to the most appropriate "winning" class. This paradigm must be followed for all the classification tasks.

3 Real Data Set

The Glass Identification data set[1] 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.

[1] 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/.

4 Bayesian and Anti-Bayesian Training and Testing

With regard to Bayesian training and testing, do the following:

1. Perform a Bayesian classification assuming that all the random variables are *independent*.
2. Perform a Bayesian classification assuming that all the random variables are *fully dependent* based on their covariance matrix.
3. Perform an Anti-Bayesian classification by using the methods taught in class, and by taking a majority vote on the decisions made on the individual features.

5 Binary-valued *Artificial/Real* Data Sets

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 DepT structure between the various features. This DepT 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.
5. To get *binary* features for the real-life data (i.e., for training and classifying using the DT and DepT), adopt a thresholding mechanism.

6 Binary-valued Training and Testing

With regard to training and testing (again, use a 5-fold cross-validation), do the following:

1. 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.
2. Using estimates of the $v_{i,j}$'s, estimate the true but unknown DepT. Record the results of how good your estimate of the true but unknown DepT is.
3. Perform a Bayesian classification assuming that all the random variables are *independent*. Here, 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 DepT that you have inferred.

Also perform all the DT and DepT classification tasks for the real-life data set rendered binary.

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

7 Report

1. Write a 8-10 page report summarizing all your results. The report should be relatively formal.

2. Compare the classification accuracy of the Bayesian and Anti-Bayesian classification methods that you have obtained for the artificial and real-life data sets.

3. Compare the classification accuracy of the DT and DepT you have obtained for the artificial and real-life data sets.

4. Compare the classification accuracy of the four algorithms ((a) Bayes, (b) Anti-Bayes, (c) using DTs, and (d) using DepTs) for the *artificial* data sets. Do some seem to outperform others? Discuss the possible reasons for these results.

5. Compare the classification accuracy of the four algorithms ((a) Bayes, (b) Anti-Bayes, (c) using DTs, and (d) using DepTs) for the real-life data sets. Do some seem to outperform others? Again, discuss the possible reasons for these results.