COMP 4106 - Artificial Intelligence
Winter 2016

Assignment #3

Due date: March 31, 2016

Machine Learning with Bayesian Classification and Decision Trees

1 Introduction

In this assignment you will be implementing a few classification algorithms including the one for Decision Trees (DTs) and using them to classify several different data sets. Please note that:

- The details of how one obtains the classifiers for the Gaussian distribution are posted on the web.
- The data sets are available on the course website as Datasets.zip.
- Each of the 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.

2 Techniques to be Implemented

The methods you will implement are:

1. The Bayesian optimal classifier, `Optimal_Bayesian`, (OB), where the distributions are assumed to be multi-dimensional Normal (Gaussian). In this case, you must assume that each class has its own mean, M_i and covariance matrix, Σ_i. If we ignore the term which contains $Determinant(\Sigma_i)$, the optimal classifier assigns the unknown testing sample, X, to the class which minimizes the Mahalanobis distance $(X - M_i)^T \Sigma_i^{-1} (X - M_i)$. Notice that Σ_i will not, in general, be diagonal. You can assume that your classification algorithm learns the values of the parameters, M_i and Σ_i from the training data sets in any way you want.
2. The Naive Bayesian classifier, `Naive_Bayesian`, (NB), where the distributions are assumed to be multi-dimensional Normal (Gaussian). In this case, we assume that the features are independent, and all the classes have *diagonal* covariance matrices. Again, if we ignore the term which contains $Determinant(\Sigma_i)$, the optimal classifier assigns the unknown testing sample X to the class which minimizes the so-called Mahalanobis distance $(X - M_i)^T \Sigma_i^{-1} (X - M_i)$. Notice that since we assume that Σ_i is diagonal, the computations are much easier. Again, you can obtain the estimates of the parameters, M_i and Σ_i, in any way you prefer.
3. The Linear Bayesian classifier, `Linear_Bayesian`, (LB), where the distributions are assumed to be multi-dimensional Normal (Gaussian). In this case, we assume that the features are independent, and all the classes have *same* covariance matrices. Here, the classifier is linear, as shown in class, and all you do in testing is to assign the testing sample, X, to the class whose mean is the closest in the Euclidean sense. Again, you can obtain the estimates of the parameters in any way you prefer.
4. The Decision Tree algorithm, `Decision_Tree`, (DT), taught in class.

3 Data Sets

There are three data sets located on the course website. We list them below.

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. Class: 1 - No presence; 2 - Cond. 1; 3 - Cond. 2; 4 - Cond. 3; 5 - Cond. 4

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

In all the above cases, ignore all the features that are non-numeric.

4 Techniques to be Implemented

1. Use a 10-fold cross-validation scheme for training and testing for the methods below. Then re-do the same with the *Leave-One-Out* method and compare their accuracies.

2. Perform all the tasks given in Section 2 on these real-life data sets. 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

3. For the DT algorithm, have your program output the resulting DT. The output[1] should be neatly indented for easy viewing. With respect to the DT algorithm, briefly discuss the causes of overfitting and how it might affect the classification accuracy.

5 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 four algorithms for the real-life data sets. Do some seem to outperform others? Discuss the possible reasons for these results.

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