Data Mining
Information Systems Department
Faculty of Computers and Artificial Intelligence
Cairo University

Assignment 3

Classification

Instructions:

1. Assignment should be done individually; copies or any other method of cheating will be graded to - 5.
2. Each student can solve one problem or both, and in the second case we will consider the higher mark.
3. Total grade is 5 marks.
4. No late submissions are allowed.
5. Don't use any built-in library. You should implement the code by yourself.
6. Discussion will be during the office hours of Eng. Esraa Salah and Eng. Doaa Galeb.
7. Deadline will be on 5/5 until 11:55 pm
8. Your program should include a user friendly interface.
9. The interface should enable user to select the percentage of the data needed to be read from the input file e.g. if the file contains 100 records, and the user needs to read 70% of the file then the analysis should be done on 70 records only.
10. The program should enable the user to select the file needed to be analyzed.
11. Using the programming language, you prefer, write a program with the following specifications:
    a. **Inputs**:
        i. A file with a set of transactions (Excel, text, etc…). (Hint. The file attached).
        ii. Divide the data set into 2 subsets, 1st one will be 75% of the data and call it "Training Set", 2nd set will be 25% of the data, and call it" Testing set"
        iii. The size of each data set will be determined as an input from the user.
        iv. The Class Label column will be the last column in the file chosen.
    b. Outputs
        i. The accuracy of the model.
        ii. The class labels for the data records provided by the user.

Problem 1

Description:

- You will be given a dataset for **diabetes dataset**, the data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This data used to predict diabetes in patients based on their medical history and demographic information.

- For this assignment you are being asked to apply the **Bayesian classifier** and **decision tree** classifier to identifying patients who may be at risk of developing diabetes.

- The class label is diabetes which is the last column in the provided comma separated file.

Requirements:

1. Apply the Bayesian and Decision Tree algorithms to build two classifier model, using the "Training set".
2. Then apply the two classifier models on the "Testing data set" to calculate the accuracy of the classifiers.
3. Compare the results of the 2 classifiers: Bayesian and Decision tree.

Problem 2

Description:

- You will be given a dataset for bank loans.
- The data set includes 5000 observations with fourteen variables divided into four different measurement categories. The binary category has five variables, including the target variable personal loan, securities account, CD account, online banking, and credit card. The interval category contains five variables: age, experience, income, CC avg, and mortgage. The ordinal category includes the variables family and education. The last category is nominal with an ID and zip code. The variable ID does not add any interesting information, e.g., an individual association between a person (indicated by ID), and the loan does not provide any general conclusion for future potential loan customers.
- In this assignment, the main goal is to classify potential customers who are more likely to purchase a loan.
    - Hint: use the most related attributes that will affect the prediction result. (See dataset description)
- For this assignment, you are being asked to apply the **Backpropagation Neural Network** and the **k-Nearest-Neighbor** Classifiers to correctly access whether the customer accepted the personal loan offered in the last campaign or not, which is the target variable in this dataset.
- The class label is y: has the customer accepted the personal loan or not (binary: 'yes', 'no')

Requirements:

1. Apply the Backpropagation Neural Network and the k-Nearest-Neighbor algorithms to build a two-classifier model using the "training set".
2. Then apply the two classifier models to the "testing data set" to calculate the accuracy of the classifiers.
3. Compare the results of the two classifiers: the Backpropagation Neural Network and the k-Nearest-Neighbor.

Data Description:

- **ID:** Customer ID
- **Age:** Customer's age in completed years
- **Experience:** #years of professional experience
- **Income:** Annual income of the customer (in thousand dollars)
- **ZIP Code:** Home Address ZIP code.
- **Family:** the family size of the customer
- **CCAvg:** Average spending on credit cards per month (in thousand dollars)
- **Education:** Education Level. 1: Undergrad; 2: Graduate;3: Advanced/Professional
- **Mortgage:** value of the house mortgage, if any. (in thousand dollars)
- **Securities_Account:** Does the customer have a securities account with the bank?
- **CD_Account:** Does the customer have a certificate of deposit (CD) account with the bank?

- **Online:** Do customers use internet banking facilities?
- **CreditCard:** Does the customer use a credit card issued by any other bank (excluding All Life Bank)?
- **Personal_Loan:** represents whether the customer accepted the personal loan offered in the last campaign or not, which is the target variable in this dataset.

Important Notes:

- As a preprocessing step, clean your data if it has noise values, e.g., if the dataset contains negative values for experience. Considering that the values of this feature indicate work experience in years, these negative values are considered noise. We assume that these values are incorrectly recorded as negative, so try to replace them with their absolute value before you start modeling.
- If you have any features that have a high correlation (>0.98) drop them, as it will affect in your model, correlated features degrade the learning performance and causes instability on the models.