BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI K. K. BIRLA GOA CAMPUS

First Semester 2018-19

Data Mining (CS F415)

Assignment-2

Dataset: The classification goal is to predict if the client will subscribe to a term deposit or not.

Answer the following questions (should be in id_number_notebook.pdf only):

A.
   a. Apply the following Classification Algorithms
      i. Naive Bayes
      ii. Decision Tree
      iii. Random Forest Algorithm with information gain criterion.
   b. Show the following
      i. Accuracy
      ii. F-Score
      iii. Recall
      iv. Confusion Matrix (Actual numbers)
      v. AUC Graph

B. Implement Bagging with Decision Tree Classifier
   a. set depth limit of the classifier as 5
   b. set depth limit of the classifier as 40

Insight Questions (should be in id_number_report.pdf only)

Note: Explain in the context of given practical.

1. Do you need data pre-processing? If Yes, mention all the pre-processing steps required and why? Else Why not?
2. How Optimal Depth Limit for Decision Tree was chosen and why do you think it is the best possible?
3. Which algorithm (mentioned in que A (a)) will be chosen based on given metrics? Is Accuracy measure alone enough to decide which model is best? Why or Why not?
4. What is the significance of AUC graph? Explain how AUC graph is useful to decide the admissibility of a model for the given problem.
5. Compare the two models (mentioned in que B) in terms of underfitting and overfitting. Explain.

Attribute List:
1 - age (numeric)
2 - job : type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3 - marital : marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4 - education (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5 - default: has credit in default? (categorical: 'no','yes','unknown')
6 - housing: has housing loan? (categorical: 'no','yes','unknown')
7 - loan: has personal loan? (categorical: 'no','yes','unknown')
# related with the last contact of the current campaign:
8 - contact: contact communication type (categorical: 'cellular','telephone')
9 - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10 - day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11 - duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no'). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model.
# other attributes:
12 - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
13 - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
14 - previous: number of contacts performed before this campaign and for this client (numeric)
15 - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
# social and economic context attributes

Output variable (desired target):
16 - y - has the client subscribed a term deposit? (binary: 'yes','no')