Problem Set 8, Part I

Important: Your work for Part II will go in a separate file. See the assignment for more details.

Problem 1: Selecting an approach

1-1)

Numeric Estimation

The approach is data mining because we attempt to find the relationship between various characteristics of a publicly traded company (input attributes) and the probability of splitting its stock (output attribute). The probability value is a numeric output attribute, so we use numeric estimation.

1-2)

Association Learning

The approach is data mining because we attempt to explore the relationships among multiple attributes. It does not focus on predicting a particular attribute, so we use association learning.

1-3)

Database Queries

The approach is database queries because we attempt to extract the total number of customers at the bank who have a money market account, which is factual information.

1-4)

Classification Learning

The approach is data mining because we attempt to extract relationships. Whether or not an individual is a good candidate for a home mortgage loan is a nominal output attribute that we want to predict, so we use classification learning.

Problem 2: Interpreting a confusion matrix

2-1)

(40 + 42 + 4) / (40 + 5 + 3 + 2 + 42 + 1 + 3 + 0 + 4) = 86 / 100 = 86%

2-2)

1 - 86% = 14%

2-3)

Actual Dem: 40 + 5 + 3 = 48

Actual Rep: 2 + 42 + 1 = 45

Actual Ind: 3 + 0 + 4 = 7

48 Democrats, 45 Republicans, and 7 Independents are actually in the Senate.

2-4)

2 actual Republicans were classified as belonging to the Democratic party.

0 actual Independents were classified as Republicans.

2-5)

40 / (40 + 2 + 3) = 40 / 45 = 88.89%

Problem 3: Comparing models

The overall accuracy of both two models = 170 / 200 = 85%

Model A is better at classifying actual accept, while Model B is better at classifying actual reject. Therefore, the validity of the model depends on the proportion of actual acceptance and actual rejection of home mortgage applications.