Data Mining Fall 2018

Project 3: Classification

Due: see Canvas
Points: 100

Please submit your report in PDF format and code (if necessary) in a separate file.

Introduction

For this project we will look again at hospitalization data (hospital3.zip). The basic idea is to classify patients depending on how much time they spend in a hospital. You have data for several years Y1-Y3 and you would like to use data of one year (e.g., Y1) to predict which patients will spend a long time in the hospital in the following year (e.g., Y2).

Follow the CRISP-DM Framework for your Report

3. Data Preparation [35 points]

- Define your classes (e.g., 0 days is "no stay", 1-5 is "short stay" and >5 is "long stay"). Explain why you defined the classes this way (maybe you want to look at the data first).
- Combine files as needed to prepare the data set for classification. You will need a single table with a class attribute to learn a model.
- Select features, create additional features, and deal with missing data.

4. Modeling [45 points]

- Create at least 3 different classification models (different techniques) and discuss the advantages of each model for this classification task.
- Assess how well each model performs (use training/test data, cross validation, etc. as appropriate).

5. Evaluation [5 points]

- How useful is your model for the health care industry? How would you measure the model's value if it was used.

6. Deployment [5]

- How would your model be used by the health care industry. How often would the model be updated, etc.

Exceptional Work [10 points]