IE 5400 Healthcare Systems Modeling and Analysis
Assignment 03
Spring 2021

Instructor: *Chun-An Chou*

Due 5 PM on 03/11/2021

Question 1 (60%)

a. Read the paper regarding Caesarian Section prediction. The dataset of 80 samples, posted on Canvas, includes five numerical and categorical variables to represent patient conditions in two classes (C-section = Yes and C-section = No).

b. Implement decision tree (DT) classifier in a 20:80 validation in Python. Compare your results with the baseline shown in Table 2 in the paper. You need to present the same information. You may try some 'tricks' (e.g., prune the DT) to improve your accuracy balancing between training and testing subsets. Note that the result shown in the paper is based on the whole dataset as the training dataset.

c. Apply association rule mining (Apriori algorithm) to find strong association rules (compared to the results on page 39 in Lecturenote 09) . You need to determine settings $minsup$ and $minconf$ by yourself. Note that some features are numerical, so you decide a reasonable approach to convert them into a categorical format to ensure 'strong' rules with 'high' support and 'high' confidence.

d. Compare the resulting rules and associated accuracy from Part (b) and Part (c). Explain your findings from the comparison. They may or may not be similar or the same to each other.

e. Implement a logistical regression analysis for the whole dataset (as the training set) and explain the feature importance correlating to diagnosis outcome?

f. Implement SVM for the whole dataset (as the training set) and explain the model with support vectors?

Question 2 (10%)

The coronary artery bypass grafting (CABG) database of the Providence Health System is analyzed using logistic regression to predict the risk of death. Two important variables are selected in analysis: patient age at operation (AGE) and a history of acute or chronic renal insufficiency (RENAL). AGE is a continuous variable measured in years and RENAL is a dichotomous variable coded as either 0 (absent) or 1 (present). RENAL is defined as a history of acute or chronic renal insufficiency or a history of a serum creatinine 2.0 recorded in the clinical record. The analysis result is shown in the following table. (Resource: Understanding Logistic Regression Analysis in Clinical Reports: An Introduction, Richard P. Anderson, Ruyun Jin, and Gary L. Grunkemeier, 2003)

(a) What is the logistical regression model?

(b) What are the odds ratios or relative importance of each independent variable in determining the outcome?

	Coeff.	Standard Error	p Value	Odds Ratio	CI Lower Limit	CI Upper Limit
AGE	0.073	0.006	<0.001	1.076	1.062	1.090
RENAL	1.162	0.177	<0.001	3.198	2.259	4.526
Constant	−8.868	0.471	<0.001			

AGE = age in years; RENAL = history of renal insufficiency; Coeff. = coefficient expressed in logits; CI = 95% confidence interval for the odds ratio.

(c) For 50-year-old patient with RENAL status, what is the probability of death?

Question 3 (20%)

Construct a Decision Tree model for the Weather Forecast dataset on pages 7-12 in Lecturenote 09. You are required to compute information gain for choosing variables and splits.

Question 4 (10%)

Given a dataset of 10 patients, there are two patient variables collected. We aim to develop a SVM classifier for disease prediction (y = 1 for disease and y = -1 for non-disease). The data and SVM results are provided in the table.

(a) Write the exact SVM model.

(b) Predict the outcome of a new patient (3, 4) using this SVM model.

i	x_{i1}	x_{i2}	y_i	α_i
$\mathbf{x}_1$	4	2.9	1	0.414
$\mathbf{x}_2$	4	4	1	0
$\mathbf{x}_3$	1	2.5	−1	0
$\mathbf{x}_4$	2.5	1	−1	0.018
$\mathbf{x}_5$	4.9	4.5	1	0
$\mathbf{x}_6$	1.9	1.9	−1	0
$\mathbf{x}_7$	3.5	4	1	0.018
$\mathbf{x}_8$	0.5	1.5	−1	0
$\mathbf{x}_9$	2	2.1	−1	0.414
$\mathbf{x}_{10}$	4.5	2.5	1	0