2022-1 CSI4117 Data Mining: Individual Project II

Due Date: 2022. 05.04 (by 11:59 pm)
Contact: prixt@yonsei.ac.kr

Submission

Upload your report at the 'Data Mining Project 2' in the LearnUS course site. Submit a single zip file (DM-P2-2021000000-maryjane.zip) containing the report (DM-P2-2021000000-maryjane.pdf) and activity annotation file. Each student will be assigned three people to annotate. Late submission penalty is taking 10% off the total score for every day.

Dataset: Available on the LearnUS course site.
In addition to the dataset for project #1, additional user_information.csv file is provided, which is "depression score" and "depression class" for each person. "depression score" represents depression level of users in the range of 0~1, and "depression class" represents one of the five levels of depression. Partition the data into training (60%) and validation (40%) sets.

Objectives of the project

This project aims to exercise the prediction and classification models with the preprocessed data from project I. It requires an understanding of how each model works with different hyperparameters. Furthermore, for the exercise of k-nearest neighbor classifier, you need to annotate the lifelog data with appropriate daily activities.

Problem

Q1. Prediction modeling with multiple linear regression.

a. Fit a multiple linear regression model to the depression score (DS) as a function of *tag_id*, *step*, and *battery_low*. Write the equation for predicting the depression score from the predictors in the model. (Note: You need to convert the categorical *tag_id* into dummy variables to use in regression models.)
b. Using the estimated regression model, what depression score is predicted for user 495 and 496? What is the prediction error?
c. Use stepwise regression with the three options (backward, forward, both) to reduce the remaining predictors as follows: Run stepwise on the training set. Choose the top model from each stepwise run. Then use each of these models separately to predict the validation set. Compare RMSE, MAPE, and mean error, as well as lift charts. Finally, describe the best model.

Q2. Classification modeling with naïve Bayes classifier.

a. Run a naïve Bayes classifier on the training set with the relevant predictors (and the depression class (DC) as the response). Note that all predictors should be categorical. Show the confusion matrix.
b. Compute the overall error, specificity, sensitivity, false-positive error, and false-negative error for the validation set.
c. Examine the conditional probabilities of the output. What is the probability of a user having 'moderately severe' depression when their average total steps per day is below 10000?
$P(DC = "Moderately\ severe" | (average\ total\ steps\ per\ day) \leq 10000)$?

Q3. Classification modeling with k-nearest neighbors classifier for daily activities. First, you need to annotate the daily activities from the 13 actions for the three people assigned as follows with the provided tagging tool.

Sleep (2)	Food (3)	Hygiene (4)	Wellbeing (3)	Culture (1)
Sleep, Wake up	Eat, Drink, Cook	Clean, Go to bathroom, Take shower, Wash dishes	Take medicine, Go walk, Work out	Watch TV

a. Following the instruction in "How to Create Annotations (Q3).pdf", annotate the activities for the three people designated to you (User1~3), and describe the life patterns for each person.
b. Fit a k-nearest neighbor classifier using your annotated data for User1 and User2. Then classify the activities of User3 using the best k. Explain the daily activities of User3, and discuss the effect of k.