Chair of Transportation Systems Engineering
Department of Civil, Geo and Environmental Engineering
Technical University of Munich

Transportation Systems M.Sc.	Statistical Learning and Data Analytics for Transportation Systems
Problem set 3	Due Date: July 27, 2020, 15:00

Classification, Clustering and PCA

Problem 1 (40 points)

(a) The file *ps3-1.csv* contains a data set with 34 features ($X_1, X_2, \ldots X_{34}$) and 1 target variable (Y). Estimate a classifier model using Support Vector Machine and Random Forest algorithm in R, respectively, with the first 14628 rows of the data. Optimize your model so that a False Positive Rate is less than 10% for Y = 0 (actual Y = 1 cases falsely classified as Y = 0).

Particularly, you are required to review relevant literature, use the k-fold cross-validation method to train the Random Forest model. Use grid search to find hyper-parameter setting: the best number of trees and features, maximum leaf nodes. Assess importance of each feature based on two criteria: Mean Decrease Accuracy and Mean Decrease Gini.

Compare two estimation methods with confusion matrices.

(b) Use two estimated models to predict the last 200 rows of the data and compare prediction with the observed target value (Y).

The report must include description of the model estimation, the R code, and the results.

Problem 2 (30 points)

The file *ps3-2.csv* contains a data set with the coordinates in degrees (longitude/latitude) of the start and end points of the trips. Each row represents a trip. Use a clustering method in R to divide the start and end points into clusters respectively. The criterion used for clustering is that the maximum distance between the points in each cluster is less than 0.03. Treat a cluster of start points as an origin for a trip, and a cluster of end points as a destination for a trip. Construct an O-D matrix to indicate the number of trips between origins and destinations. Your report must include description of the approach used for clustering, the code, and the results.

Problem 3 (30 points)

The file *ps3-1.csv* contains a data set with 34 features to predict a target Y. Please use principal component analysis in R to reduce the dimensionality of the features. Determine the optimum number of principal components using a criterion that 95 % of total variance can be explained. Your report must include description of the steps of principal component analysis, the code, and the results.