2018 Spring, INE5008 Data Mining, by Kichun Lee

HW 1

Due 1 pm, April 6, 2017

**PART #1**

Refer to the attached *data\_1.txt*. You may also refer to MS Excel or a R package, *plot3D*:

http://cran.r-project.org/web/packages/plot3D/vignettes/plot3D.pdf

Draw a 3-dimensional plot with X coordinate ‘KM’, Y coordinate ‘Weight’, and Z coordinate ‘Price’.

**PART #2**

Refer to the attached *data\_1.txt*. We are considering the following three regression models for Price:

(Model 1) Price =

(Model 2) Price =

(Model 3) Price =

Use (1) 5-fold cross-validation and (2) randomized 80%training-20%testing validation, and find the best model among the three.

**PART #3**

Refer to the following R scripts and the url:

https://cran.r-project.org/web/packages/titanic/titanic.pdf

#install.packages("titanic")

library("titanic")

data(titanic\_train)

str(titanic\_train)

titanic\_train\_original <- titanic\_train

dropNames <- c("Name", "Ticket", "Cabin", "PassengerId")

toSelect <- !(names(titanic\_train\_original) %in% dropNames)

titanic\_train2 <- titanic\_train\_original[, toSelect]

Reading the description in titanic.pdf and using data-frame *titanic\_train2* in the R script, find categorical variables in the data-frame. Convert them into numerical variables by one-hot-encoding (, i.e., creating a dummy variable for each level).

(Open question) As we saw in class exploratory data analysis (EDA) as in the following url and class material, perform your own EDA:

<https://www.kaggle.com/notaapple/detailed-exploratory-data-analysis-using-r>

EDAreferenceCode.zip

The EDA should include analysis of missing values, histogram of each variable, boxplot of numerical variables depending on a categorical variable, if any, scatter plot, correlation heat-map, and so on.