15.071: The Analytics Edge **Fall 2022**

Homework 2: Nonlinear Regression and Regularization

Out: September 26; Due: October 9th, 11:59pm.
Please post the assignment in pdf format with file name "Lastname_15071_HW2.pdf".
For each question, please include the main `R` *commands that you used in your submission.*

Problem 1: Predicting Pollution Levels in Boston Suburbs [50 pts]

In this problem, we'll be working with a dataset on quality of life in Boston suburbs and try to predict pollution levels using a weighted mean distance from major employment centers in various suburbs of Boston. Below is a brief description of the variables of interest in this problem:

- `nox` Nitrogen oxides concentration (parts per 10 million).
- `dis` Weighted mean of distances to five Boston employment centers.
- `nonretail` Concentration of non-retail businesses. (= 1 if High; 0 otherwise).

We are interested in understanding the dependence of nitrous oxide concentration (as a proxy for pollution levels) on a weighted mean distance from the top five employment centers in Boston. To do this, first read `Boston.csv` into `R` and split the data into training and test set using stratified sampling as follows:

```
set.seed(123)
split = createDataPartition(boston$nox, p = 0.7, list = FALSE)
boston.train = boston[split,]
boston.test = boston[-split,]
```

a) Train a linear regression model to predict nitrous oxide concentration `nox` using the weighted mean distance from the top five employment centers in Boston (`dis`) as a predictor. Report R^2 and out-of-sample OSR^2 of your model. Plot your regression line along with the data. **[10 pts]**

b) Now, train a cubic polynomial regression model (degree=3) using `dis` as an independent variable. Evaluate and report R^2 on your training and test datasets. Visualize your cubic polynomial regression model. **[10 pts]**

c) Create a scatter plot of your training data by plotting `nox` vs. `dis`, and use `nonretail` variable for color gradient in order to segregate suburbs with high and low concentration of non-retail businesses.

Now, train a new cubic polynomial regression model by incorporating interaction terms between `dis` and `nonretail`. Report both R^2 and OSR^2 of your new model. In the figure created above, plot this new cubic polynomial model. **[15 pts]**

d) We'll now fit a cubic spline model to predict the concentration of nitrous oxide `nox`. Recall that the total degrees of freedom of a spline model depends on the degree of the polynomial ($m = 3$, for cubic splines) and the number of knots (K). Place knots at 20th, 40th, 60th and 80th percentiles of `dis` column. To achieve this, first find these quantiles using the following command:

```
knots <- quantile(boston.train$dis, p = c(0.2, 0.4, 0.6, 0.8))
```

Now, train a natural cubic spline model using interaction terms between `dis` and `nonretail`, and knots at the locations as discussed above. Compare this model with the linear regression model and the polynomial regression model with and without interactions terms in terms of OSR^2. **[15 pts]**

Problem 2: Predicting Housing Prices in King County [50 pts]

In this problem we consider the King County housing prices dataset, which describes sales of 432 properties in King County, Washington. You will work with the dataset provided in the `kc_house.csv` file. This dataset contains 1 dependent variable (price) and 17 independent variables, as listed below:

- bedrooms: The number of bedrooms in the house.
- bathrooms: The number of bathrooms in the house.
- sqft_living: The square footage of the living space.
- sqft_lot: The square footage of the lot.
- floors: The number of floors in the house.
- waterfront: Whether this house has a view to waterfront.
- view: Times it has been viewed.
- condition: How good the condition is (overall).
- grade: Overall grade given to the housing unit, based on King County grading system.
- sqft_above: The square footage of the house apart from basement.
- yr_built: The year when the house is built.
- yr_renovated: The year when the house is renovated.
- zipcode: The zip code of the house.
- lat: Latitude coordinate.
- long: Longitude coordinate.
- sqft_living15: The square footage of interior living space of the nearest 15 neighbours.
- sqft_lot15: The square footage of the land lots of the nearest 15 neighbours.

First, import the dataset:

```
kc_raw <- read.csv("kc_house.csv")
```

a) Explore the dataset using the following steps.

   i) Report the correlation matrix between the predictors. What do you observe? [**5 pts**]

   ii) Plot the price as function of each one of 2-3 predictors that you consider important. How does a change in each predictor affect the salary? [**5 pts**]

b) Feature normalization is often important in regularized regression problems. Normalize the data, and split them into a training set and a test set, using the following commands (please use the same seed):

```
pp <- preProcess(kc_raw, method=c("center", "scale"))
kc <- predict(pp, kc_raw)
set.seed(123)
train.obs <- sort(sample(seq_len(nrow(kc)), 0.7*nrow(kc)))
train <- kc[train.obs,]
test <- kc[-train.obs,]
```

The first models you will try are simple linear regression models.

i) Fit a linear regression model with all the predictors using the training set, and make predictions on the test set. Report the in-sample and out-of-sample R^2. Comment briefly on the sign and significance of the variables and the R^2 values. Does this make sense, in view of your earlier observations? **[10 pts]**

ii) Fit a restricted linear regression model by manually selecting only the predictors that are significant in the full model. (We will consider a variable significant only if the p-value is below 0.1) Report the in-sample and out-of-sample R^2. **[5 pts]**

c) Let's now examine if regularization can help us do better.

i) Train ridge regression and LASSO models with 5-fold cross-validation to select the appropriate value of the shrinkage parameter λ, using the Mean Squared Error as the performance metric (use `cv.glmnet()` function). Plot the cross-validated Mean Squared Error as a function of λ. Report the value of λ that minimizes the Mean Squared Error for each method. **[10 pts]**

ii) With the selected values of λ, re-train your ridge regression and LASSO models on the full training set. Report each model's coefficients and comment on the effects of ridge regression vs. LASSO. Use each model to make predictions on the test set. Report the values of the in-sample R^2 and the out-of-sample R^2. **[10 pts]**

iii) As we discussed in class, LASSO often results in models where many coefficients are set to 0. Is this the case with the LASSO model you trained in question c) part *ii*)? **[5 pts]**