# Homework 7

Due Nov 12 at 3:00.

**Searching for solutions is forbidden.**

**(1)** Download the Ames Housing data:

```
library(AmesHousing)  ###you need to install this package first
ames = make_ames()
names(ames)
X = ames[,-79]  ##covariates
Y = ames[,79]   ##the reponse
X = Filter(is.numeric,X)  ##get rid of factor variables for simplicity
dim(X)
names(X)
str(X)
Y = unlist(Y)    ##make Y into a vector
X = as.matrix(X) ##make X into a matrix
```

The goal is to predict sale price. Divide the data into two groups: training data and test data. Make the test data have size 500. For example:

```
n = nrow(X)
I = sample(1:n,size=500,replace=FALSE)
Xtest = X[I,]
Ytest = Y[I]
Xtrain = X[-I,]
Ytrain = Y[-I]
```

(a) Using the first 10 covariates only of the training data (i.e. the first 10 columns of $Xtrain$) fit an additive model. Plot the fitted functions and summarize your findings.

(b) Fit a random forest using the first 10 covariates only of the training data. Summarize the variable importance measure and compare to the additive model. Do they agree on which variables are important?

(c) Use the test data to compare the prediction accuracy of the forest and the additive model. Provide 95 percent confidence intervals for the prediction accuracy in each case and a 95 percent confidence interval for the difference.

(d) Fit a random forest using all covariates in $X$. Estimate the prediction accuracy on the test data. Does this predict better than using only 10 covariates? (Again, make sure you get the estimate and the confidence interval.)

(e) Consider $Y = \log(\text{Sale\_Price})$ and $X = \log(\text{Lot\_Area})$. Use nonparametric quantile regression to find the 10th percentile, the median, and the 90th pecentile of $Y$ given $X$. Plot the data and add your estimated quantiles.

**(2)** Suppose that $X \in \mathbb{R}$ and $Y \in \{1, \ldots, k\}$. Assume that, the distribution of $X$ given $Y = j$ is $N(\mu_j, \sigma^2)$. Let $h_*(x)$ be the optimal classifier that minimizes $P(Y \neq h_*(X))$. Find an explicit expression for $h_*(x)$.

**(3)** Install the package ISLR then get the Auto data:

```
library(ISLR)
attach(Auto)
names(Auto)
str(auto)
```

Create a new variable $Y$ that is 1 if `mpg` is above its median and 0 if `mpg` is below its median.

(a) Explore the data graphically. Which features seem most relevant for predicting $Y$?

(b) Randomly split the data into a training set and a test set. You can decide how big each of these should be.

(c) Using the training data, build two classifiers using: (i) logistic regression, (ii) a random forest.

(d) Use the test data to estimate the predictive accuracy of your classifiers. (Provide a point estimate and confidence interval for the error rate of each classifier.) Which method works best?