ISE-5970: Energy Analytics

Homework 5

Due: Sunday November 17th, 11:59 p.m.

Before start, please read the following.

1. The questions in this homework allow you to practice your R skills for ARIMA models (Chapter 8) and machine learning algorithms (LR, KNN, and SVM).
2. For all questions, you must submit **1) the source file that contains the R commands, and 2) the snapshot of what R outputs after you run your R program.**
3. **There will be 30% penalty for late submissions per day.**
4. I strongly prefer if you **electronically submit** your homework through Canvas by putting all files in a zip folder.
5. Please assign numbers for each solutions, so it would be easy for me to read the answers.

Good Luck! ☺

Question 1 (5+5+5+10+5+5+5=40 credits):

Consider usmelec, the total net generation of electricity (in billion kilowatt hours) by the U.S. electric industry (monthly for the period January 1973 – June 2013). In general, there are two peaks per year: in mid-summer and mid-winter.

a. Examine the 12-month moving average of this series to see what kind of trend is involved.
b. Do the data need transforming? If so, find a suitable transformation.
c. Are the data stationary? If not, find an appropriate differencing which yields stationary data.
d. Identify a couple of ARIMA models that might be useful in describing the time series. Which of your models is the best according to their AIC values?
e. Estimate the parameters of your best model and do diagnostic testing on the residuals. Do the residuals resemble white noise? If not, try to find another ARIMA model which fits better.
f. Forecast the next 15 years of electricity generation by the U.S. electric industry. Get the latest figures from the EIA (https://bit.ly/usmelec) to check the accuracy of your forecasts.
g. Eventually, the prediction intervals are so wide that the forecasts are not particularly useful. How many years of forecasts do you think are sufficiently accurate to be usable?

Question 2 (5+5+5+5+5+5= 30 credits):

For the mcopper data:

a. If necessary, find a suitable Box-Cox transformation for the data;
b. Fit a suitable ARIMA model to the transformed data using auto.arima() ;
c. Try some other plausible models by experimenting with the orders chosen;
d. Choose what you think is the best model and check the residual diagnostics;
e. Produce forecasts of your fitted model. Do the forecasts look reasonable?
f. Compare the results with what you would obtain using ets() (with no transformation).

Question 3 (10+8+7+5 = 30 credits):

This question should be answered using the Weekly data set, which exists in the ISLR package. This data is similar in nature to the Smarket data from the chapter’s lab, except that it contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

a. Fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held-out data (that is, the data from 2009 and 2010).
b. Repeat (a) using KNN with K=5.
c. Now, change K in KNN and obtain confusion matrix and accuracy measures. What is the best value for K?
d. Repeat (a) using Linear SVM.