CEE 154/254 Data Analytics for Physical Systems
Autumn 2020
Assignment 3

Assignment 3 is due on 11/2/2020 at the beginning of class

Goal: Apply various regression techniques to real-world data, understand impact of over-fitting and benefits of regularization. Covers lectures 8-10.

For this assignment we will use the Foshan static sensor data.

Hint: to have a consistent numeric input to the models and make the analysis easier, you can convert the time vector to the timestamp in second with the following MATLAB commands. For this homework, you will use this timestamp as input to the polynomial fitting, linear regression, and AR models.

```
load hw3_1.mat
pm2d5= data.pm2d5;
time = data.time;

time_second = (datenum(time)-floor(datenum(time)))*24*60*60;
```

Part I [30 points]: **Linear Regression.** Consider the time series data for Foshan PM2.5. Load the data file *hw3_1.mat*. The suggested MATLAB commands for this part of assignment is *polyfit(__)* and *polyval(__)*.

a) Determine the linear regression fit and 95% confidence interval for the time series data. Plot the time series data and overlay the linear fit line and confidence bounds on the same plot.
b) Determine a polynomial regression fit with degree of 3 and its 95% confidence interval. Plot the time series data and overlay the polynomial fit and confidence bounds. Describe the differences between this plot and the plot in a).
c) Determine the polynomial regression fit with degree of i) 5; ii) 7; and iii) 50 along with their 95% confidence bounds. Plot each fit in a separate window and discuss the differences.
d) Which of the polynomial fits do you believe best represents the data, and why? Describe the benefits of higher order fits as well as the drawbacks.

Part II [30 points]: **Regularization + Other Factors**. Consider the time series data for Foshan PM2.5. Load the data file *hw3_1.mat*. Create a 3 x 2 grid of subplots as described below.

a) In the first subplot window, plot the polynomial regression fit and time series data with a 20-degree polynomial.

b) High degree of polynomial curve fitting tends to overfit the data, so we need to balance between bias and variance. One way to overcome the problem of overfitting is called ridge regression. What is the cost function of ridge regression? Make sure you clearly define all the variables in the equation. Does it have a closed-form solution for polynomial curve fitting? If so, write it down for 20-degree polynomial curve fitting.

c) Another way to overcome overfitting is LASSO regression. What is the cost function of LASSO regression? How is it different from the cost function of ridge regression? Does it have a closed-form solution for polynomial curve fitting?

d) Pick the best values for the regularization parameter $\lambda \in \{0.1, 0.01, 0.001\}$ for ridge and LASSO regressions. Then, in the second and third subplot windows, plot the fit using ridge regression and LASSO regression fits using a 20-degree polynomial. What λ value leads to the best performing ridge regression? What about for LASSO? Describe the differences between these plots and the plot from a). How are the results from ridge regression different from those obtained from LASSO? Describe both in reference to the plot itself as well as the coefficients obtained.
The suggested MATLAB command for Lasso regression is *lasso(   )*.

e) In the final three subplot windows, plot the regression fit using LOWESS with kernel span (number of data points for calculating the smoothing value) of i) 5, ii) 100, and iii) 1000. Describe the differences between each of these three subplots and comment on which kernel span/bandwidth you think best represents the data, and why.
The suggested MATLAB command for LOWESS is *smooth(__,'lowess')*. This function uses the traditional tri-cube kernel function that weights data points based on the distance between them to the center. However, any kernel functions (e.g., RBF kernel) could also be used.

Extra points (5 points): implement LOWESS for polynomial fitting on your own using the RBF kernel. Set the RBF kernel parameter (σ) to i) 100, ii) 2000, and iii) 20000 and redo e) plot.

f) Create a 2 x 2 subplot. In the first window plot the linear regression fit and the Foshan time series data. In the remaining subplots, plot the linear regression fit with additional features of: i) Time + Humidity; ii) Time + Humidity + Speed; and iii) Time + Humidity + Speed + Temperature. Describe how the additional features influence the linear regression fit. Do any of these features improve the fit? Justify your answer.

Part III [30 points]: **Autoregressive model.** Consider the time series data for Foshan PM2.5. First, load the data file *hw3_3.mat.* This file contains a 101-by-2 matrix that includes timestamp (the first column) and displacement (the second column) of a 2-degree-of-freedom oscillator with harmonic force (a sinusoid function) applied.

a) Fit the first 80 displacement measurements with a 5-order autoregressive model AR(5) and predict the last 20 displacement measurements. Create a 2-by-1grid of subplots. In the first subplot, plot the 80 observed displacement records, your predictions, and the ground truth of the last 20 measurements, and the 95% confidence interval. In the second subplot, plot the residual, which is the difference between ground truth and the prediction. Calculate the mean squared error between your prediction and the ground truth.

b) Fit the first 80 displacement measurements with a 10-order autoregressive model AR(10) and predict the last 20 displacement measurements. Create a similar plot as a) using your new AR(10). Calculate the mean squared error between your prediction and the ground truth. Describe the difference between this plot and the plot in a). Which of the AR models do you think provides a better prediction? Why?

Then, consider the time series data for Foshan PM2.5. Load the data file *hw3_1.mat.*

c) Fit the first 20-hour PM 2.5 measurements with a 10-order autoregressive model AR(10) and predict the last 4 hours PM 2.5 concentration. Create a similar plot as a) using your new AR(30). Calculate the mean squared error between your prediction and the ground truth. Describe the difference between this prediction result/plot and the result/plot in b). Do you think increasing the order of the AR model would help on improving the prediction performance? Do you think the AR model is a suitable method for predicting PM 2.5? Why or why not?