CEE 154/254 Data Analytics for Physical Systems
Autumn 2023
Assignment 4

Assignment 4 is due on 11/29/2023 midnight

Goal: Explore autoregressive modeling, Gaussian process regression, sampling bias. Understand potential pitfalls of extrapolation and overfitting in regression models. Covers lectures 9-10.

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

Part I [40 points]: **Cross-Validation, Extrapolation, and Bias.** Consider the time series data for Foshan PM2.5 collected every minute. Load the data file *hw4_1.mat.*

a) Determine the polynomial regression fit (degree = 3) on the time series data from 0:00am – 9:59pm using a 5-fold cross-validation (CV). Randomly partition the dataset into 5 parts and choose one partition as test data and the rest as training data for each validation. Plot all 5 fits on one figure with the entire time series data. (10 points)

b) Using each regression model from a), predict the PM2.5 value from 10:00pm – 11:59pm. Plot each prediction curve on the same figure as a). (4 points)

c) Repeat a) and b) above using polynomial regression with degrees i) 5 and ii) 9. Describe the differences in the extrapolated (predicted) PM2.5 values. Discuss the dangers of model overfitting and pitfalls of extrapolation. (8 points)

d) Determine the polynomial regression fit (degree = 5) using the entire time series data and plot the fit along with the raw data. Randomly remove 70% of the time series data whose values are smaller than the median PM2.5 value from the dataset. Determine the polynomial regression fit (degree 5) with this reduced dataset. Plot the resulting regression fit on the same plot. Describe how the fit is affected by the dataset bias. (8 points)

e) Randomly remove 35% of the time series data (from the entire dataset). Determine the polynomial regression fit (degree = 5) with this reduced dataset. Plot the resulting regression fit with the results from d) above. Describe how the fit is different from the biased dataset fit in d) and how it compares to the fit on the entire dataset. (6 points)

f) Discuss how biased datasets can influence regression fits and describe possible strategies for reducing the effect of bias. (4 points)

Part II [25 points]: **Autoregressive model.** Load the data file *hw4_2_1.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 70 displacement measurements with a 5-order autoregressive model AR(5) and predict the last 30 displacement measurements. Create a 2-by-1grid of subplots. In the first

subplot, plot the 70 observed displacement records, your predictions, and the ground truth of the last 30 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. (5 points)

b) Fit the first 70 displacement measurements with a 10-order autoregressive model AR(10) and predict the last 30 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? (10 points)

Then, consider the time series data for Foshan PM2.5. Load the data file *hw4_2_2.mat.*

c) Fit the first 20-hour PM 2.5 measurements with both AR(15) and AR(40). Predict the last 4 hours of PM 2.5 concentration. Create similar plots as a) and b). Calculate the mean squared error between your prediction and the ground truth for both AR(15) and AR(40). Describe the difference between these two models. 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? (10 points)

Part III [35 points]: **Gaussian Process Regression**. Consider the time series data for Foshan PM2.5 across 10 devices. Load the data file *hw4_3.mat.* Note that the spatial and temporal data has been normalized in the given code and saved as x and y for consistent analysis and visualization. You may use them directly.

a) Use an exponential kernel to estimate the GPR of the spatial-temporal PM2.5 data with a 5-fold CV and report the average RMSE error across each CV fold. For each validation, randomly choose 80% data points from the entire data to train your GPR model and validate on the rest 20% data. (8 points)

b) Create a 6 x 2 subplot. Interpolate the PM2.5 data in the spatial domain for each hour between 10:00 and 21:59 (12 plots corresponding to the 12 hours at 10am, 11am, …, 9pm) and plot the ensuing temporal-spatial PM2.5 data using a contour plot. (MATLAB function `surf()`). Describe any interesting observations from each of the plots. How do the PM2.5 values change over time and space? (12 points)
Hint: to make the contour plot, first create a 2D spatial grids (latitude and longitude) using MATLAB function `meshgrid()`, construct the feature matrix with 3 columns where each column corresponds to 0-1 normalized time, latitude, and longitude, respectively, and input this feature matrix to the trained GPR model for prediction.

c) Repeat a) – b) above, but with a squared exponential kernel. (6 points)

d) Repeat a) – b) above, but with a Matern 3/2 kernel function. (6 points)

e) How does the selected kernel affect the GPR interpolation results? Describe any notable differences between the results from each kernel. (3 points)