Assignment 3

Artificial Datasets

<u>Univariate Case</u> :

a) Generate 500 real number for the variable X from the uniform distribution U [0,1].
b) Construct the training set T = { $(x_1,y_1),(x_2,y_2),\ldots\ldots,(x_{500},y_{500})$} using the relation $Y_i = \sin(2\pi x_i) + \epsilon_i$, where $\epsilon_i \sim N(0,0.25)$.
c) In a similar way construct a testing set of size 50 .
   Test = { $(x'_1,y'_1),(x'_2,y'_2),\ldots\ldots,(x'_{50},y'_{50})$}
d) Estimate the regularized least squared polynomial regression model of order k = 1,2, 3, 7, using the training set T. For example,
   (i) For k=1 , we need to estimate $F(x) = w_1x + b$
   (ii) For k = 2, $F(x) = w_2x^2 + w_1x + b$.

e) List the value of coefficients of estimated regularized least squared polynomial regression models for each case.
f) Obtain the prediction on testing set and compute the RMSE for regularized least squared polynomial regression models for order k=1,2,3 and 7.
g) Plot the estimate obtained by regularized least squared polynomial regression models for order k =1,2,3 and 7 for training set along with y_1, y_2, .., ,y_{20} . Also plot our actual mean estimate $E(Y/x_i) = \sin(2\pi x_i)$.
h) Plot the estimate obtained by regularized least squared polynomial regression models for order k =1,2,3 and 7 for testing set along with y'_1, y'_2,.. , y'_{50}. Also plot the $\sin(2\pi x'_i)$. Study the effect of regularization parameter λ on testing RMSE and flexibility of curve and list your observations.

Bivariate Case :

a) Construct the training set T = {(x_1,y_1),(x_2,y_2),……,(x_{200},y_{200})} using the relation $Y_i = \sin(2\pi(||\mathbf{x_i}||) + \epsilon_i$ where $\epsilon_i \sim N(0,0.25)$ and $\mathbf{x_i} = [x_{i1}, x_{i2}]$, where x_{i1} and x_{i2} are from U[0,1]. In the similar way construct a testing set of size 50

Test = { (x'_1,y'_1),(x'_2,y'_2),……,(x'_{50},y'_{50})} .

b) Obtain the prediction on testing set and compute the RMSE for regularized least squared polynomial regression models for order k =1,2 and 3. Also plot the estimated function and target function for the training set and testing set.

Real-world Datasets

a) Consider the motorcycle dataset. Estimate the Regularized Least Square regression models using the M sigmoidal basis functions. The k^{th} sigmoidal basis function can be obtained using

$$\sigma(w_k, b_k, x) = \frac{1}{1 + e^{-(w_k{}^T x + b_k)}}$$

$$where\ , w_k\ is\ any\ vector\ of\ R^n\ \ and\ b_k\ is\ a\ real\ number$$

I. Plot the estimated function and obtain the training RMSE error for M = 2, 5, 10. What happens when you increase the number of basis functions.

II. For M =10, find the minimum mean and standard deviations of RMSE using leave-one out method by tuning the parameter λ.

III Estimate the Regularized Least Square kernel regression model using Gaussian kernel which is given by

$$k(x,z) = e^{-\frac{||x-z||^2}{2\sigma^2}}$$, for any two points x and $z \in R^n$.

Also, find the minimum mean and standard deviations of RMSE using leave-one out method by tuning the parameter λ and kernel parameter σ .