**Math 6350 fall 2019 MSDS Robert Azencott**

**Final Exam (Take Home) is a project which must be done individually by each Math 6350 student**

**Due date Mon Dec 9th at Midnight**

**Kernel Ridge Regression applied to one real data set**

*Question 1 : Download a real data set DS*

DS must include N cases X(1)... X(n) with n moderately large (800 < n <1500)

each case X(j) can be viewed as vector in Rp described by

p features X1(j)... Xp(j) which are the "explanatory variables"

each case X(j) is associated to an observed value Yj of the target variable Y

each feature must be a "continuous" variable;

avoid or eliminate discrete features taking only a small number of values ;

make sure that p ≥ 10 in your Data Set and make sure to include an artificial feature Xp(j) =1 for all cases j=1...n

the main technical goal is to apply Kernel Ridge Regression (KRR) to predict the value of Y whenever a new case X = [ X1... Xp ] is given

Describe your data set as concretely as possible

Explain the practical impact of obtaining good predictions of Y when the explanatory variables are known

for each feature X1 X2 ... Xp, compute and display its mean and standard deviation

same question for Y

Split the data set DS into a training set TRAIN and a test set TEST, with respective proportions 80% , 20%

Compute the empirical correlations cor(X1, Y) ... cor(Xp,Y) and their absolute values C1 ... Cp

compute the 3 largest values among C1 ... Cp, to be denoted Cu > Cv > Cw which are

display separately the 3 scatter plots (Xu(j), Yj) , (Xv(j), Yj) , (Xw(j), Yj) where j= 1...n

Interpret the results

*Question 2: Kernel Ridge Regression (KRR) with radial kernel*

*For this question we use intensively the training set TRAIN which has size m = 80% n*

The m cases in TRAIN are (after reordering of their indices) denoted X(1)... X(m) to simplify notations . Select the kernel = "radial "kernel K(x,y) defined for x and y in Rp by the formula

K(x,y) = exp(-gamma || x- y ||**2**) where gamma >0 is a parameter to be selected later

the KRR approach involves also a cost parameter 1/λ which roughly evaluates the cost of a prediction error. The parameter λ >0 will also have to be selected later

*Once "* λ *" and "gamma" are selected , the best KRR prediction function pred(x) is defined for any input vector x in Rp by the formula*

pred(x) = y (G + λ Id)**-1** V(x)

where

y= [Y1 ...Ym] is a line vector

Id = mxm identity matrix

V(x) is a *column* vector with m coordinates V1(x) , ... , Vm(x) given by Vj(x) = K(x, X(j))

the mxm matrix G is the kernel gramian G= [ Gij ] with G(i,j) = K(X(i), X(j)) for all i ,j in [1...m]

Compute the matrix G and its eigenvalues L1 >L2 > ... > Lm ≥ 0

Plot Lj versus j

Plot the increasing ratios RATj= (L1 + ... + Lj)/(L1+ ... + Lm)

Identify the smallest j such that RATj ≥ 95% and set λ = Lj

Select at random two lists List1 and List 2 of 100 random integers each , within [1...m]

For all i in List 1 and all j in List2 compute Dij = ||X(i) -X(j)||

Plot the histogram of the 10000 numbers Dij

Compute q =10% quantile of the 10000 numbers Dij

Set gamma = 1/q

Compute the matrix M = G + λ Id and its inverse M**-1**

As seen in class the prediction formula becomes

pred(x) = A1 K(x, X(1)) + ... + Am K(x,X(m))

compute the line vector A = [A1 ... Am] by A= y M**-1**

Compute the RMSEtrain of the prediction function pred(x) by running it on all x in TRAIN set

Compute the RMSEtest of the prediction function pred(x) by running it on all x in TEST set

Compare these two RMSE values , and compute their ratios RMSE/ avy where

avy = mean of the m absolute values |Y1|, ... , |Ym|

Interpret the results

*Question 3: Improving the results through step by step tuning*

Repeat the preceding operations for other pairs of parameters gamma and λ

Suggestion: change only one parameter at a time to check in which direction to go for improved performances

Select the best choice of parameters in terms of accuracy RMSE/avy and stability of performance when one goes from TRAIN to TEST set

Identify the 10 cases in the TEST set for which the squared prediction error is the largest

Vizualise the 10 cases by performing a PCA analysis and projecting all the TEST cases onto the first 3 principal eigenvectors of the PCA correlation matrix

Try to identify what went wrong with the prediction of the absolute worse case X(w) by looking at the terms involved in pred(Xw) and comparing to another case where the prediction erro is really small

*Question 4 : Analysis of the best predicting formula pred(x)*

Fix the best choice of parameters as found in the preceding question.

reorder the |A1|, |A2|, ....|Am| in decreasing order , which gives a list B1 > B2 ... > Bm >0 and plot the decreasing curve Bj versus j

Compute the ratios bj = (B1 + ... + Bj)/(B1 + ...+Bm)and plot the increasing curve bj versus j

Compute the smaller j such that bj > 99%. and the corresponding threshold value THR = Bj

For i =1... m, if |Ai|> THR set AAi = Ai and otherwise set AAi = 0. This yields a reduced formula

PRED(x) = AA1 K(x, X(1)) + ... + AAm K(x,X(m))

Run this reduced formula on the TRAIN and TEST sets to evaluate its performances

Compare these performances to the original formula pred(x) and interpret the results

*Question 5 (optional): Implement KRR using a pre existing function*

using the best parameters found above try to use a pre-existing software function implementing the KRR technique

<https://archive.ics.uci.edu/ml/datasets/Electrical+Grid+Stability+Simulated+Data+>