Spring 2019: CSCI 6990 Programming Assignment #1

DUE: Monday, March 11, 2019 (**Softcopy** @4 PM; **Hardcopy**@6:30 PM/in class)

Instructions

- **All work must be your own other than the instructor provided data/code and hints to be used. You are NOT to work in teams on this assignment.**

Problem Description:

Forest Fire Prediction: we want to develop model(s) for predicting forest fire from given datasets.

Forest-fire or bushfire is a serious environmental issue, creates economic and ecological damages and threatens human lives in many places in the world such as Arizona, Australia, Argentina, Canada, New Zealand, Portugal etc. Fast detection or, prediction can help firefighters greatly and reduce casualties. Further, if the size of the affected area can be predicted at the same time, it will provide an estimation of resources required to fight the fire.

The possible major resources-categories to collect data to build-up our model can be: (a) satellite-based datasets, (b) setup remote-sensors and collect sensor datasets, (c) collect meteorological datasets. Since automatic meteorological stations are often available and the dataset is available at a lower cost - we will develop a model based on the meteorological dataset (see Moodle for the dataset) [1].

Data Description:

The given data has 517 instances – each data point is 12 dimensional and the output column (13^{th} column) is the burned area given in hectors (ha). The features or attributes including the output (burned) "area" are:

SL	Features	Comments
1	X	x-axis spatial coordinate of a park map: 1 to 9
2	Y	y-axis spatial coordinate of a park map: 1 to 9
3	Month	month of the year: 'jan' to 'dec'
4	Day	day of the week: 'mon' to 'sun'
5	FFMC	Fine Fuel Moisture Code [2], ranges: 18.7 to 96.20
6	DMC	Duff Moisture Code, ranges: 1.1 to 291.3
7	DC	Drought Code, ranges: 7.9 to 860.6
8	ISI	Initial Spread Index, ranges: 0.0 to 56.10
9	temp	temperature in Celsius degrees: 2.2 to 33.30
10	RH	relative humidity in %: 15.0 to 100
11	wind	wind speed in km/h: 0.40 to 9.40
12	rain	rain in mm/m^2: 0.0 to 6.4
13	area	the burned area of the forest (in ha): 0.00 to 1090.84

PART (A)

Training Dataset: the file name 'data' contains the dataset, that you will use. Test data is not separately provided; therefore, you will apply 10-fold cross-validation (FCV).

Rank the Features: Rank the features by computing the correlation between *i* and *j*, where *i* is any one of the features from 1 to 12, and *j* =13th feature. Computed all the 12 correlation values and place them in **Table #1** [Feature id (1 to 12), correlation value] in descending order of the absolute value of the correlation, in your report. Plot the correlation graphs (**Graph #1 to 5**) for the top 5 input features.

Model Prediction: You will need to predict the models or learners using non-iterative equation, B=inv(X'*X)*X'*Y, for models of order M=1, 2, 4, 6, and a value of *your-choice* (YC) using (a) all the 12 input features and (b) top 5 input features from **Table #1**.

Performance Evaluation and Error Calculation: We will evaluate the performance of each the model by computing mean-absolute-error (MAE) and root-mean-squared-error (RMSE). MAE and RMSE are computed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 / N}$$

For each Model (M=1..6, YC), **Table #2** will have M=*i*, average (avg) and standard deviation (SD) of MAE of the 10FCV; as well as avg., and SD of RMSE of the 10FCV in each row – in the experiment all the 12 input feature will be included. Include **Table #2** in your report. Now, use the top 5 input features for your experiment and similarly generate **Table #3**.

Plots: Plot **Graph#6** and **Graph#7** for: (a) Models (M=1..6, YC) versus MAE, (b) Models (M=1..6, YC) versus RMSE where the models were trained using 12 input features. Similarly, provide **Graph#8** and **Graph#9** for models those are built using top 5 input features only.

PART (B)

Regularization: For a model of order M=6 and YC, you need to apply the regularization using the non-iterative equation: $B = (\mathbf{X}^T\mathbf{X} + \lambda M_\lambda)^{-1}\mathbf{X}^T\mathbf{y}$, where M_λ is the identity matrix with $M_\lambda(1, 1) = 0$, and *lambda* (λ) is the regularization parameter. Compute the model error for, λ= (a) 0 (which you have already computed in PART (A)), (b) 0.5E-8, (c) 1.5E-6, (d) 2.0E-4, (e) 1 and (f) 2. Compute the *Errors* as we have done in PART(A) and plot for "λ versus the MAE" and "λ versus R-MSE" for both 12 input-features as well as 5 top input-features based models as we have done in PART(A).

Important Note: you may need to apply *ln λ* along the *x*-axis for plotting as well as you may need to do the same along the *y*-axis.

Submission: Submit a report where you will briefly describe your experiment. Further, the report must contain all the **Tables** and **Graphs**. Put the report, your generated dataset and code in a compressed-folder and upload it via Moodle.

Hardcopy: please provide the hardcopy of the report only.

References:

[1] P. Cortez and A. Morais. (2007). *Forest Fires Dataset, http://www3.dsi.uminho.pt/pcortez/forestfires/*.

[2] N. R. Canada. (2008). *Canadian Wildland Fire Information System, http://cwfis.cfs.nrcan.gc.ca/background/summary/fwi*.

--- x ---