Salinity - (Assignment 4 - Question 2)

- subject Machine Learning / AI

DESCRIPTION

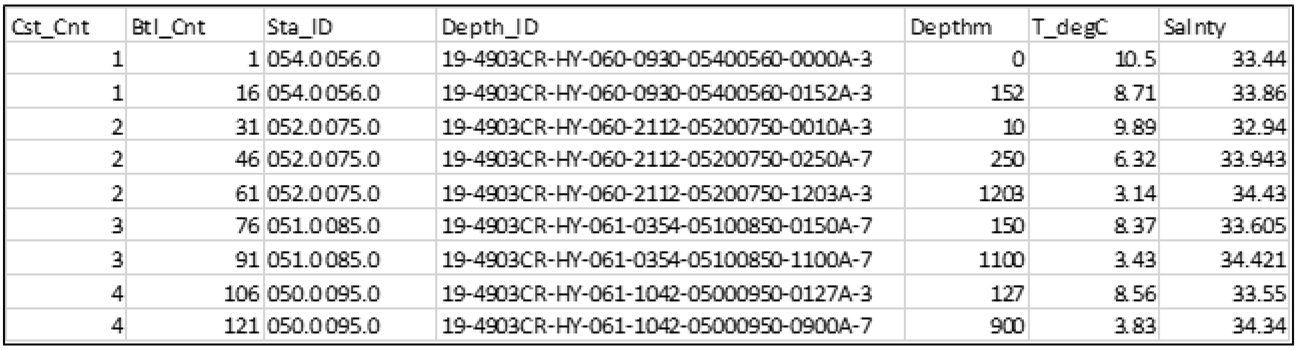
The California Cooperative Oceanic Fisheries (CalCOFI) are a unique partnership of the California Department of Fish & Wildlife, NOAA Fisheries Service and Scripps Institution of Oceanography.

CalCOFI conducts quarterly cruises off southern & central California, collecting a suite of hydrographic and biological data on station and underway. Data collected at depths down to 500 m include elements such as: temperature, salinity, oxygen, phosphate, silicate, nitrate and nitrite, chlorophyll and a few other elements.

Here's a brief description of the columns of the input data set:

- **Cst_Cnt**: Auto-numbered Cast Count – Cumulatively increasing count with each station (Cast equivalent to a single survey)
- **Btl_Cnt**: Auto-numbered Bottle count – Cumulatively increasing count (several bottles obtained in each cast)
- **Sta_ID**: CalCOFI designated Stations' IDs
- **Depth_ID**: [Century]-[YY][MM][ShipCode]-[CastType][Julian Day]-[CastTime]-[Line][Sta][Depth][Bottle]-[Rec_Ind]
- **Depthm**: Depth in meters at which the samples are taken
- **T_degC**: Water temperature in degree Celsius
- **Salnty**: Salinity in g of salt per kg of water (g/kg)

This is a preview of the data under consideration:



Cst_Cnt	Btl_Cnt	Sta_ID	Depth_ID	Depthm	T_degC	Salnty
1	1	054.0 056.0	19-4903CR-HY-060-0930-05400560-0000A-3	0	10.5	33.44
1	16	054.0 056.0	19-4903CR-HY-060-0930-05400560-0152A-3	152	8.71	33.86
2	31	052.0 075.0	19-4903CR-HY-060-2112-05200750-0010A-3	10	9.89	32.94
2	46	052.0 075.0	19-4903CR-HY-060-2112-05200750-0250A-7	250	6.32	33.943
2	61	052.0 075.0	19-4903CR-HY-060-2112-05200750-1203A-3	1203	3.14	34.43
3	76	051.0 085.0	19-4903CR-HY-061-0354-05100850-0150A-7	150	8.37	33.605
3	91	051.0 085.0	19-4903CR-HY-061-0354-05100850-1100A-7	1100	3.43	34.421
4	106	050.0 095.0	19-4903CR-HY-061-1042-05000950-0127A-3	127	8.56	33.55
4	121	050.0 095.0	19-4903CR-HY-061-1042-05000950-0900A-7	900	3.83	34.34

As part of this exercise, we are interested in finding the relationship between water salinity & water temperature. Use the input data file **bottle.csv** to read the above data set.

Perform the analysis as described in the following steps using Python-

Preparing the data:

Load 1000 rows of the data set from the file **bottle.csv**. You will only need the columns **Salnty** and **T_degC** for performing the below analysis. Replace the null values with their respective means.

Questions:

1. Split the data set into train and test data with 30% for testing with random_state=1. Use **Salnty** as **X** and **T_degC** as **Y.** *Print* the number of rows in the train and test data sets. (**Hint**: Use function from sklearn library with default random state)

2. Perform simple linear regression on the training data set. Calculate and *print* the **coefficient** & **intercept** of the simple linear regression model created using the training data.

(**Hint:** Use function from sklearn library. Employ the **metrics** to calculate **Mean squared Error** and **R2 score**. These values will be used in question no. **3** below.)

3. Predict the values for the test data set and measure the score based on the predicted values as below:

- Using the linear regression model created in question **II** above**,** measure-
- **Mean Squared Error (MSE)**
- **R^2 score**
- *Print* these two values

4. Perform cross validation using k-fold as required below:

- Use number of folds equal to 5 and random_state=5
- Generate Root Mean Squared Error (RMSE) scores , 1 for each of the 5 iterations and calculate its mean
- *Print* the value of its **mean RMSE** score
- **Hint**: Use method **cross_validate**

******NOTE: While using the K-Fold, pass the entire dataset by only splitting into independent and dependent features i.e X, y. Do not use train test split into this.***

Input Format:

- You have to read data from a file named **bottle.csv** present at the location **/data/training/bottle.csv**

Output Format:

- You have to perform the operations as required by the above questions and write (written above as **print**) your output to **2** separate files named **output1.csv** and **output2.csv** both of which should be present at the location **/code/output/<filename>.csv**
- **output1.csv** should contain 4 rows of data- The **number of rows of the train and test data** calculated in **step 1** should be written on the first and the second row respectively in a form such as: **500** and **400.**

  ******(First row should contain the no of rows of train data and second row of the output file should contain the no of rows of test data.)***

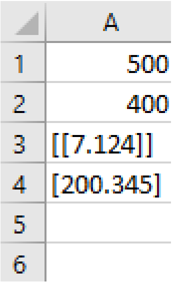
- The third and the fourth row should contain the values of coefficient and intercept respectively, both rounded to 3 decimal places in a form such as: **[[1.234]]** and **[100.123]**
- **output2.csv** should contain 3 rows of data - the values of **MSE** and **R^2 score,** both rounded to 3 decimal places on the first and the second row respectively in a form such as: **2.234** and **5.678**
- The third row of **output2.csv** should contain the value of **mean RMSE** rounded to 3 decimal places such as **6.789**
- Do not write any headers or additional labels in the **output1.csv** and **output2.csv** file

Sample Input

Read the input file **/data/training/bottle.csv**

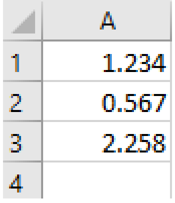
Sample Output

Example: output1.csv will have data looking like this:



	A
1	500
2	400
3	[[7.124]]
4	[200.345]
5	
6	

Example: output2.csv will have data looking like this:



	A
1	1.234
2	0.567
3	2.258
4	

DATASETS

- Training dataset