ULaval IFT & GLO

Mini-Project # 1 (on 20%)

to return on June 24, 2020 at 11.55pm

Professor Brahim Chaib-draa *GLO-7050: Machine learning in practice*

The projects of the course are the projects of the versions of the course COMP 551, course given by the colleagues of McGill, a big thank you to them for giving us the authorization to use them.

Preamble

- This mini-project is an individual work. However, this does not prevent you from chatting with other students taking the course. In no case, however, should you use the code and writing of others; you are asked to prepare your own.

- If you "borrow" ideas, methods, approaches or others, please indicate your sources in the report.

- It is strongly suggested that you argue and / or justify your answers.
- After the submission date, you have up to a week to submit your work with a 30% penalty.

- You are free to use libraries such as Numpy or Scipy for Python. However, you should not use pre-existing implementations of the requested algorithms, you should implement them by yourself.

- if you have any questions regarding the work, please go through the Forum, asking your questions clearly.

1 Model selection

For this experiment, you must use the Dataset-1 containing files for training, validation and testing. Both the input and the output at this dataset are real scalars. The dataset is generated from a degree polynomial *not* and a slight Gaussian noise is added to the target.

- Adjust the data to a polynomial of degree 20. Then report the MSE (Mean-Square Error) for training and validation. Do not use any accruals. View the model-data adjustment and comment on it.

- Add now a regularization L 2 to your model. Knowing that λ varies from 0 at
  1 then plot for different values of λ the MSE for training and validation. Then choose the best value of λ and then give the performance-test for the model considered. Visualize the model-data fit and comment on it. item What do you think is the degree of the polynomial? Can it be inferred from the visualization of the previous question?

2 Gradient descent for regression

For this experiment, you must use Dataset-2 containing files for training, validation and testing. The input and output at the level of this dataset are both scalars with real value.

- Set up a linear regression model for this dataset using stochastic gradient descent. You must use the online-SGD (Stochastic Gradient Descente) (ie, SGD-with one example at a time). Use a step size of

  1 *e* - 6. Calculate the MSE for the validation set and this for each epoch.
- Try different step sizes and choose the best one using the validation file. Then give the MSE for the test.
- Visualize the regression adjustment for each epoch by giving 5 visualizations which show how the regression adjustment evolves during the training process.

3 Dataset from “real life”

For this question, you must use the “Communities and Crime” data set which can be extracted from UCI: (http://archive.ics.uci.edu/ml/datasets

- The dataset is from real life and as such it probably would not have the "good" properties that one would expect. Therefore, your first job is to make the dataset *usable* by completing all the omitted values. For each of the missing attributes, use the sample mean from each column. Is it a good choice ? What else could you use? If you have a better method, describe it and use it to complete the omitted data.

- The dataset now completed, then set up a regression model. Report the 5-part cross-validation error to know: report the MSE (mean squared error) of the best fit, data-model, performed on the test data, averaged over 5 different sharing 80-20 data; as well as the parameters learned for each of the 5 models.

- Now use Ridge regression with the dataset above. Repeat the experiment for different values of λ and report the MSE for each of the values, on the test data, averaged over 5 different data sharing 80-20; as well as the parameters learned. Which of the values of λ gives the best fit? Is it possible to use the information obtained during this experiment to select characteristics? If so, what is the best fit you could achieve with a reduced feature set?

Sharing data 80-20: Some suggestions

1. Make 5 different 80-20 splits in the data and name them as *CandC-train <num> .csv* and *CandC-test <num> .csv.*

2. For all 5 datasets that you have generated, learn a regression model using the 80% data and test it using 20% data.

3. Report the average MSE over these 5 different runs.

Instructions for submitting the code

1. Please submit a simple zipped folder with your Name.
2. Using Python, the submitted solution should be of the Jupyter Notebook type.
3. Make sure that all the files allowing your code to roll, are supplied with the "correct" access paths. Your code should be rolled over without any modification.

Instructions for submitting the report

1. Your report should not be verbose, just the basics. When asked for comments, you should not go beyond 3 to 4 lines for each of these comments.

2. Report all visualizations as far as possible (curves, graphs, adjustments, etc.).

3. Either you use colors to separate the different graphs, or you use different symbols.