INF8245E(Fall 2021) : Machine Learning | Given on: Sep 13, 10:00 pm

Assignment #1 | Due on : Oct 04, 10:00 am

- This is an individual assignment. However, you are allowed to discuss the problems with other students in the class. But you should write your own code and report.
- If you have any discussion with others, you should acknowledge the discussion in your report by mentioning their name.
- You have to submit the `pdf` copy of the report on gradescope before the deadline. If you handwrite your solutions, you need to scan the pages, merge them to a single `pdf` file and submit. Mark page 1 for the outline item *'Verbosity'* on gradescope.
- Submit all the text files and codes compressed to a single file `<your-matricule>.zip` on **Moodle**.
  Note: gradescope doesn't accept .zip.
- Be precise with your explanations in the report. Unnecessary verbosity will be penalized.
- You are free to use libraries with general utilities, such as matplotlib, numpy and scipy for python. However, you should implement the algorithms yourself, which means **you should not use pre-existing implementations of the algorithms as found in SciKit learn, Tensorflow, etc.**!
- If you have questions regarding the assignment, you can ask for clarifications in Piazza.

1 Sampling

1. A grad student's daily routine is defined as a multinomial distribution, p, over the set of following activities:
   - Movies: 0.2
   - INF8245E: 0.4
   - Playing: 0.1
   - Studying: 0.3

   1. Every morning, the student wakes up and randomly samples from this distribution an activity to do for the rest of the day. Provided that you can only sample from uniform distribution over (0,1), write a pseudocode to sample from the given multinomial distribution.
   2. Implement your sampling algorithm and use it to sample the student's routine for 100 days. Report the fraction of days spent in each activity. Now use it to sample for 1000 days. Report the fraction of days spent in each activity. Compare these fractions to the underlying multinomial distribution.

2 Model Selection

You have to use Dataset-1 for this experiment. Dataset-1 consists of train, validation, and test files. The input is a real valued scalar and the output is also a real valued scalar. The dataset is generated from an n-degree polynomial and a small Gaussian noise is added to the target.

1. Fit a 20-degree polynomial to the data.
   (a) Report the training and validation RMSE (Root Mean-Square Error). Do not use any regularization.
   (b) Visualize the fit.
   (c) Is the model overfitting or underfitting ? Why?
2. Now add L2 regularization to your model. Vary the value of λ from 0 to 1, with a 0.01 step size.
   (a) For different values of λ, plot the training RMSE and the validation RMSE.
   (b) Find the best value of λ and report the test performance for the corresponding model.
   (c) Visualize the fit for the chosen model.
   (d) Is the model overfitting or underfitting ? Why?
3. What do you think is the degree of the source polynomial? Can you infer that from the visualization produced in the previous question?

3 Gradient Descent for Regression

You have to use Dataset-2 for this experiment. Dataset-2 consists of train, validation, and test files. The input is a real valued scalar and the output is also a real valued scalar.

1. Fit a linear regression model to this dataset by using stochastic gradient descent (one example at a time).
   (a) Using a step size of 10^{-4}, plot the training and validation RMSE against the number of epochs, until convergence.
2. Try different step sizes and choose the best step size by using the validation data.
   (a) Report in a table the validation performance with different step-sizes.
   (b) Report the test RMSE of the chosen model.
3. Report 5 different visualizations chosen at random to illustrate how the regression fit evolves during the training process.

4. Repeat part 1 using full-batch gradient descent.

5. Comment on the difference between full-batch gradient descent and stochastic gradient descent based on your experiments.

4 Real life dataset

For this question, you will use the Communities and Crime Data Set from the UCI repository (`http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`).

1. This is a real-life data set and as such, it would not have the nice properties that we expect. Your first job is to make this dataset usable by filling in all the missing values.

   (a) Use the sample mean of each column to fill in the missing attributes. Is this is a good choice? Explain why or why not.

   (b) What else might you use to fill in the missing attributes?

   (c) If you have a better method, describe it, and use it for filling in the missing data. Explain why your method is better.

   (d) Turn in the completed data set.

2. Use the first 20% of the dataset for testing and use the remaining 80% for training in the order given in the dataset file.

   (a) Report the 5-fold cross-validation average RMSE.

   (b) Report the test RMSE.

3. We now use Ridge-regression on the above data.

   (a) In order to choose the best λ, plot the average RMSE using 5-fold cross validation, for various values of λ [x-axis: λ, y-axis: Average RMSE]. Explain how you chose the range of λ to explore.

   (b) Which value of λ gives the best fit?

   (c) Report the test RMSE using the value of λ you chose.

   (d) Is it possible to use the information obtained during this experiment for feature selection? If so, explain how?

   (e) Report the test RMSE of the best fit you achieve with a reduced set of features?

   (f) How different is the performance of the model with reduced features compared to the model using all the features? Comment about the difference.

Instruction for code submission

1. Submit a single zipped folder with your matricule id as the name of the folder. For example if your matricule ID is 12345678, then the submission should be 12345678.zip.
2. You can only use Python 3 and you must submit your solution as a jupyter notebook.
3. Make sure all the data files needed to run your code are within the folder and loaded with relative path. We should be able to run your code without making any modifications.

Instruction for report submission

1. You report should be brief and to the point.
2. Report all the visualizations (learning curves, regression fit).
3. Do not include your code in the report!