Coursework **Fall 2022**
Machine Learning (B555) **Roni Khardon**
Department of Computer Science, IU Bloomington

Programming Project 2

This assignment is due by Wednesday 10/12, 11:59pm via Canvas.

Programming language and libraries: You can write your code in any programming language so long as we are able to test it on SICE servers (python should be ok; please ask about other options). We plan to run some or all or submitted code for further testing and validation. You may use standard I/O, math, and plotting libraries (including numpy, and matplotlib). However, other than these, please write all the code yourself without referring to special libraries or modules, i.e., no scikit, no pandas, no other data processing libraries etc.

Academic integrity: a reminder that each student must work out solutions to problems on their own and write-out the solution and/or programs independently. Please see the course canvas page for a more detailed policy.

Overview: Experiments with Bayesian Linear Regression

Our goal in this assignment is to evaluate linear regression, its regularized variant and the Bayesian model including model selection. In all your experiments you should report the performance in terms of the mean square error

$$\text{MSE} = \frac{1}{N}\sum_i (\phi(x_i)^T w - t_i)^2$$

where the number of examples in the corresponding dataset is N.

Data

Data for this assignment is provided in a zip file `pp2data.zip` on Canvas.

Each dataset comes in 4 files with the training set in `train-name.csv` the corresponding labels (regression values) in `trainR-name.csv` and similarly for test set. For Task 1 we use the `crime` and `housing` datasets. For task 2 the files are named `f3` and `f5`; these datasets have only one feature and the label was generated from polynomial regression, using polynomials of degree 3 and 5 respectively. Note that the train/test splits are fixed and we will not change them in the assignment.

Algorithms

In this assignment you should implement two machine learning algorithms. The first algorithm is regularized linear regression, i.e., given a dataset, the solution vector w is given by equation (3.28) of [B]. Note that plugging in $\lambda = 0$ we get the maximum likelihood solution so the same code can be used for this case as well. We can then calculate the MSE on the test set using the w for prediction.

The second algorithm is the formulation of Bayesian linear regression with the simple prior $w \sim \mathcal{N}(0, \frac{1}{\alpha} I)$. Recall that the evidence function (and evidence approximation) gives a method to pick the parameters α and β. Referring to [B], the solution is given in equations (3.91), (3.92), (3.95), where m_N and S_N are given in (3.53) and (3.54). These yield an iterative algorithm for selecting α and β using the training set. This scheme is pretty stable and converges in a reasonable number of iterations. You can initialize α, β to random values in the range $[1, 10]$. We can then calculate the MSE on the test set using the MAP (m_N) for prediction. This is the same as the prediction of the regularized algorithm with $\lambda = \alpha/\beta$.

Task 1: Comparing the Bayesian algorithm to Linear Regression with and without Regularization

To evaluate the algorithms we will generate learning curves. In particular for training fractions $f \in \{0.1, 0.2, 0.3, \ldots, 1.0\}$, train the learning algorithm using the initial fraction of the dataset of that size, and calculate MSE on the test set. Each of the following should be done for both datasets.

(i) Run the model selection algorithm and report the values of α, β and effective λ for each train size.

(ii) Run the maximum likelihood algorithm and the model selection algorithm. Plot their test set MSE as a function of training set size to compare their performance. Please limit the y-axis in plots to the range [0,1] to ensure visibility of differences (you can crop MSE values or use plotting library limits). What can you observe w.r.t. their relative performance? Try to explain why the results are as observed and whether they are as expected or not.

(iii) Repeat part (ii) with values of λ equal to 1.0, 33.0, 100.0, 1000.0 and discuss the results. Can we use a single universal value for λ for different datasets? Is the Bayesian algorithm successful in selecting a good value? How could one select it otherwise?

Note: typically, learning curve experiments are repeated multiple times with randomized data to observe standard deviations in differences. In this assignment we skipped this to reduce the amount of work.

Task 2: Bayesian Model Selection for Parameters and Model Order

In this part we work with the datasets `f3` and `f5` whose labels were generated using polynomials. You should run the Bayesian model selection scheme of the previous task using polynomial degrees d in $\{1, 2, \ldots, 10\}$. The files themselves only include the x values, so in order to run the regression code you must first generate appropriate training data. For example, for degree 3, each x in the training and test files is replaced with $1, x, x^2, x^3$.

For each degree d, run the Bayesian Model Selection code to select α, β (and hence λ) and calculate the log evidence (given in eq (3.86)) on the training set. Then calculate the MSE on the test set using the MAP (m_N) for prediction. In addition, run non-regularized linear regression on the same data and calculate the MSE on the test set.

For each dataset plot the log evidence and 2 MSE values (of non-regularized and Bayesian models) as a function of d. Can the evidence be used to successfully select α, β and d for the Bayesian method? How does the non-regularized model fare in these runs?

Note: evidence is only relevant for the Bayesian method and one would need some other method to select d using maximum likelihood in this model.

Submission

Please submit two separate items via Canvas:

(1) A zip file `pp2.zip` with all your work and the report. The zip file should include: (1a) Please write a report on the experiments, include all plots and results, and your conclusions as requested above. Prepare a PDF file with this report. (1b) Your code for the assignment, including a README file that explains how to run it. When run your code should produce all the results and plots as requested above. Your code should assume that the data files will have names as specified above and will reside in sub-directory pp1data/ of the directory where the code is executed. We will read your code as part of the grading – please make sure the code is well structured and easy to follow (i.e., document it as needed). This portion can be a single file or multiple files.

(2) One PDF "printout" of all contents in 1a,1b: call this `YourName-pp2-everything.pdf`. One PDF file which includes the report, a printout of the code and the README file. We will use this file as a primary point for reading your submission and providing feedback so please include anything pertinent here.

Grading

Your assignment will be graded based on (1) the clarity of the code, (2) its correctness, (3) the presentation and discussion of the results, (4) The README file and our ability to follow the instructions and test the code.