Coursework **Fall 2021**
Machine Learning (B555) **Roni Khardon**
Department of Computer Science, IU Bloomington

Programming Project 2

This assignment is due by Wednesday 10/13, 11:59pm via Canvas.

You can write your code in any programming language so long as we are able to test it on SICE servers. We plan to run some or all or submitted code for further testing and validation.

Overview: Experiments with Bayesian Linear Regression

Your goals in this assignment are to explore the role of regularization in linear regression and to investigate two methods for model selection (evidence maximization and cross validation). 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.

We have 4 datasets and 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. We have two real datasets `crime` and `wine` and two artificial datasets `artsmall` and `artlarge`. Note that the train/test splits are fixed and we will not change them in the assignment (in order to save work and run time).

For the artificial data you can compare the MSE results to the MSE of the hidden true functions generating the data that give 0.533 (`artsmall`), and 0.557 (`artlarge`).

Task 1: Regularization

In this part we use regularized linear regression, i.e., given a dataset, the solution vector w is given by equation (3.28) of Bishop's text.

For each of the 4 datasets plot the training set MSE and the test set MSE as a function of the regularization parameter λ (use integer values in the range 0 to 150). For each dataset it is useful to put both curves on the same plot. In addition, compare these to the MSE of the true functions given above.

In your report provide the results/plots and discuss them: Why can't the training set MSE be used to select λ? How does λ affect error on the test set? Does this differ for different datasets? How do you explain these variations?

Note: The experiments in this task tell us which value of λ is best in every case *in hindsight*. That is, we need to see the test data and its labels in order to choose λ. This is clearly not a realistic setting and it does not give reliable error estimates. The next two tasks investigate methods for choosing λ automatically without using the test set.

Task 2: Model Selection using Cross Validation

In this part we use 10 fold cross validation *on the training set* to pick the value of λ in the same range as above, then retrain on the entire train set and evaluate on the test set. To avoid confusion, the procedure for doing this is explained at the end of the assignment.

Implement this scheme, apply it to the 4 datasets and report the values of λ selected, associated MSE and the run time. How do the results compare to the best test-set results from part 1 both in terms of the choice of λ and test set MSE?

Task 3: Bayesian Model Selection

In this part we consider 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 Bishop's book, 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). As discussed in class these yield an iterative algorithm for selecting α and β using the training set. We can then calculate the MSE on the test set using the MAP (m_N) for prediction.

This scheme is pretty stable and converges in a reasonable number of iterations. You can initialize α, β to random values in the range $[1, 10]$ and stop the algorithm when the difference in α, β values is < 0.0001.

Implement this scheme, apply it to the 4 datasets and report the values of α, β, the effective $\lambda = \alpha/\beta$, the associated MSE and the run time. How do the results compare to the best test-set results from part 1 both in terms of the choice of λ and test set MSE?

Task 4: Discussion of Results

Tabulate together the values obtained in tasks 1-3 and use this to discuss the following questions. How do the two model selection methods compare in terms of effective λ, test set MSE and run time? Do the results suggest conditions where one method is preferable to the other?

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, their 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 pp2data/ 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) our ability to test the code.

Addendum: 10 Fold Cross Validation for Parameter Selection (with a fixed train/test split)

We have already used cross validation for estimating accuracy in project 1. Cross validation can also be used for parameter selection if we make sure to use the train set only.

To select parameter a of algorithm A(a) over an enumerated range $a \in V_1, \ldots, V_K$ using dataset D we do the following:

1. Split the data D into 10 disjoint portions.
2. For each value of a in $V_1, \ldots, V_K$:
   (a) For each i in $1 \ldots 10$
      i. Train $A(a)$ on all portions but i and test on i recording the error on portion i
   (b) Record the average performance of a on the 10 folds.
3. Pick the value of a with the best average performance.

Now, in the above, D only includes the training set and the parameter is chosen without knowledge of the test data. We then retrain on the entire train set D using the chosen value and evaluate the result on the test set.