Coursework **Fall 2019**
Machine Learning (B555) **Roni Khardon**
Department of Computer Science, IU Bloomington

Programming Project 2

This assignment is due by Monday 10/14, 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 (i) to investigate the effect of the number of examples, the number of features, and the regularization parameter on the performance of the corresponding algorithms, and (ii) to investigate two methods for model selection in linear regression (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.

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 both artificial data and real data, with real data in the `crime` and `wine` datasets. Note that the train/test splits are fixed and we will not change them in the assignment (in order to save run time).

The files for the artificial data are named as `NumExamples-NumFeatures` for easy identification of training set characteristics. (NumExamples is the number of examples in the training set, not the test set). Specifically we have 3 variants with the combinations `100-10, 100-100, 1000-100`. Note that the different artificial datasets have different underlying predictive functions (hidden vector w) so they should not be mixed together. The artificial data was generated using the regression model and is thus useful to test the algorithms when their assumptions hold.

For the artificial data you can compare the MSE results to the MSE of the hidden true functions generating the data that give 5.714 (for 100-10), 0.533 (for 100-100), and 0.557 (on 1000-100).

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 5 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). 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?

Task 2: Learning Curves

Now pick three "representative" values of λ from the first part ("too small", "just right", and "too large") for the dataset 1000-100. For each of these values plot a learning curve for the learned regularized linear regression on this dataset.

A learning curve plots the performance (in our case MSE) of the algorithm as a function of the size of the training set. To produce these curves you will need to draw random subsets of the training set (of increasing sizes) and record the performance (on the fixed test set) when training on these subsets. To get smooth curves approximating the mean performance you will need to repeat the above several times (at least 10 times) and average the results. Use enough training set sizes between 10 and 1000 samples to generate smooth curves.

In your report provide the plots and some numerical results and discuss them: What can you observe from the plots regarding the dependence of the error on λ and on the number of samples? Consider both the case of small training set sizes and large training set sizes. How do you explain these variations?

Task 3.1: Model Selection using Cross Validation

The previous experiments 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. In this part and the next we investigate methods for choosing λ automatically without using the test set.

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 5 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 3.2: 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). 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 $[1, 10]$

Implement this scheme, apply it to the 5 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 3.3: Comparison

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? Please try to think about the results obtained and discuss these questions even if you do not see an obvious trend.

Submission

Please write clear code with sufficient documentation so that we can read it. In addition write a README file that explains how the code is organized (if in multiple files) and how to compile and run the code. If this is non-trivial please write a script that runs the code and explain how to use it in the README file. When run in this manner your code should produce all the results and plots as requested above.

For testing, we may want to run your code with different challenge data (using the same filenames as in the assignment). To facilitate this your code should assume that data files reside in the same directory as the code and should not in any way be tuned to the contents of data files provided.

Please submit two items via Canvas: (1) Please write a report on the experiments, their results, and your conclusions as requested above. Prepare a PDF file with this report. (2) Collect all your code for the assignment (including the README file) in a zip file named `pp2code.zip`. You do not need to include the data that we provided. Your code should assume that the data files will will reside the same directory where the code is executed.

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 run the code on SICE servers.

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.