

Assignment 3

This assignment is due by Thursday October 11 in class.

Overview: Experiments with Bayesian Linear Regression

On the course page you will find 3 datasets for experimentation with regularized linear regression. Each dataset comes in 4 files with the training set in `train-NumExamples-NumFeatures.csv` the corresponding labels (regression values) in `trainR-NumExamples-NumFeatures.csv` and similarly for test set. Note that the different datasets have different underlying predictive functions (hidden vector w) so they should not be mixed together. This is artificial data that I generated using the regression model and is thus useful to test the algorithms when their assumptions hold.

Note: we plan to add one or two more datasets TBA well before the deadline. This will add more runs and results to parts 1,2 and 4 and in your report but should not affect the code. Please do start the assignment early.

Your goals in this assignment are to investigate the effect of the number of examples, the number of features, and the regularization parameter on performance of the corresponding algorithms, as well as investigate the quality of techniques for model selection, i.e., for choosing the regularization parameter.

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 .

You can compare the results to the MSE of the hidden true functions generating the data that give 3.78 (for 100-10), 3.78 (for 100-100), and 4.015 (on 1000-100) on these datasets.

I recommend using matlab (or any other system with a good matrix package) but you should otherwise write the code, which should be simple enough, yourself.

Your Task

To prepare for the experiments start by creating 3 additional training sets from the training dataset 1000-100, using the initial 50, 100, and 150 examples respectively. Call these 50(1000)-100, 100(1000)-100, and 150(1000)-100. The test set does not need to be modified. This will allow us to investigate the performance more closely in some cases.

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

For each of the 6 datasets (3 original and 3 you created) plus TBA data 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). In addition compare these to the MSE of the true functions given above.

In your report provide the results/plots and discuss them: How does λ affect the result? How does the choice of λ depend on the setting of features vs. number of examples? and the number of examples when the number of features is fixed?

2. Now pick three “representative” values of λ from the first part. For each of these values plot a learning curve for the learned regularized linear regression using the dataset 1000-100.

A learning curve plots the performance 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.

In your report provide the results/plots and discuss them: what can you observe from the plots regarding the dependence on λ and the number of examples?

3. 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 cross validation to pick the regularization parameter λ . This works as follows:

- Use 10 fold cross validation *on the training set* to pick the value of λ in the same range as above. Cross validation is explained below for the benefit of those who have not seen it before.
- Once the value is chosen we train on the entire training set using this value of λ .
- We can then evaluate and calculate the MSE on the test set.

Implement this scheme and apply it to the 6 datasets plus TBA data. In your report provide the results and discuss them: How do the results compare to the choice of λ and best test set MSE in hindsight from part 1? what is the run time cost of this scheme? How does the quality depend on the number of examples and features? what are the important factors affecting performance?

4. 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 . 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.

Implement this scheme and apply it to the 6 datasets plus TBA data and report as in the previous part.

5. How do the two model selection methods compare in terms of the test set MSE and in terms of run time? what general conclusions can you make?

Submitting Your Assignment

Please submit printouts of your code, and a short report on the experiments, their results, and your conclusions from them.

Addendum: 10 Fold Cross Validation for Parameter Selection

Cross Validation is the standard method for evaluation in empirical machine learning. It 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 $a \in V_1, \dots, V_K$ 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, \dots, V_K :
 - (a) For each i in $1 \dots 10$
 - 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.