

Programming Project 2

This assignment is due by Monday, October 23 1:30pm which is class start time.
Please submit both a hardcopy (in class) and electronically (via provide).

Overview: Experiments with Bayesian Linear Regression

On the course web page you will find several datasets for experimentation with regularized linear regression. 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.

For tasks 1-3 the files for the artificial data are named as `NumExamples-NumFeatures` for easy identification of dataset characteristics (NumExamples is the number of examples in the training 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 task 4 the files are names `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.

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 performance of the corresponding algorithms, and (ii) to investigate the quality of Bayesian model selection, both for choosing the regularization parameter and for selecting model order (in polynomial regression).

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 .

For the artificial data in tasks 1-3, 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.

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 (all but **f3**, **f5**) 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: Why can't the training set be used to select λ ? How does λ affect error on the test set? How does the choice of the optimal λ vary with the number of features and number of examples? 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 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. Use enough training set sizes between 10 and 800 samples to generate smooth curves.

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 samples? Consider both the case of small training set sizes and large training set sizes. How do you explain these variations?

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). 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 and apply it to the 5 datasets. In your report provide the results and discuss them: 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? (Note that our knowledge from part 1 is with hindsight of the test set, so the question is whether model selection recovers a solution which is close to the best in hindsight.) How does the quality depend on the number of examples and features?

Task 4: 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 $\{1, 2, \dots, 10\}$. 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 . Consider selecting the degree and α, β based on log evidence. What can you conclude on the success of this method for polynomial regression? How does the non-regularized model fare in these runs? (Note that one would need cross validation to select d in this case) What can you conclude on the success of the Bayesian model when using a model that is too rich (i.e. when the degree is too high)?

Submission

- You should **submit** the following items **both electronically and in hardcopy**:
 - (1) All your source code for the assignment. Please write clear code with documentation as needed. The source code should (i) run on *homework.eecs.tufts.edu*, (ii) run from the command line *without editing* with a single command (if there is more than one execution command required, include those commands in a Bash script which we can run), and (iii) output the requested results.

You can assume the data files will be available in the same directory as where the code is executed. Please use filenames as provided for the data. Please include a short README file with the code execution command.
 - (2) A PDF report on the experiments, their results, and your conclusions as requested above.
- For electronic submission, put all the files into a zip or tar archive, for example `myfile.zip` (you do not need to submit the data we give you). Please do not use another compression format such as RAR. Then submit using `provide comp136 pp2 myfile.zip`.
- Your assignment will be graded based on the clarity and correctness of the code, and presentation and discussion of the results.