CSCI B-555: Machine Learning
Programming Project 2 Report

Aniruddha Patil (2000578987)

10/14/2019

Task 1: Regularization

- Why can't the training set MSE be used to select λ?

  We might get lucky in some situations where the train set MSE and the test set MSE are correlated. However, as seen from the graphs below, that is not the case for most of the datasets. When we select the MSE solely based on the training MSE, we do not accommodate the possibility of new data that may come up in the test set. Thus, choosing λ from training set MSE is not a good idea.

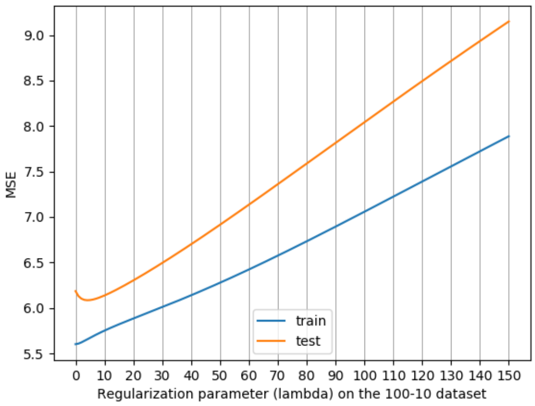
- How does λ affect error on the test set?

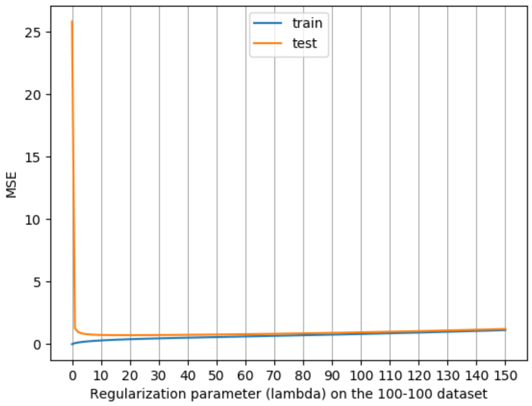
  In all datasets, from $l\lambda$ values of $0 - 150$, we can see that the MSE dips down for small values of *lambda* and comes back up as they get larger. This is an expected result as we had seen in the previous assignment. This is due to our prior being too strong which does not accommodate new data properly.

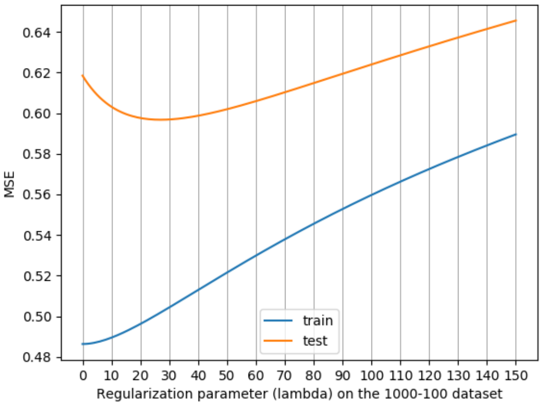
- Does this differ for different datasets?

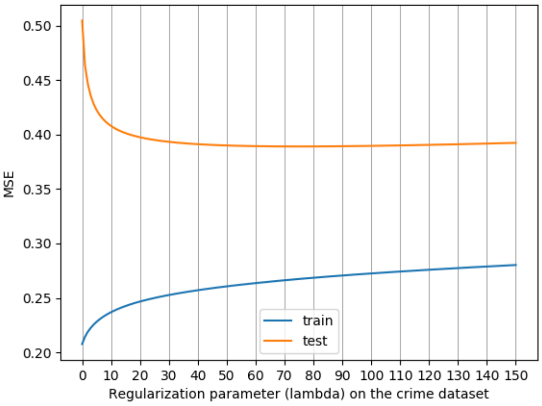
  In a general sense, as we have seen, the above trend of the prior's parameter has risen in all the datasets we have dealt with so far. So, while the MSE itself may vary, I think that the trend will stay the same for different datasets. This is assuming that their test set will always have data that is not in the train set. If this were not the case, then our prior would increase MSE and keep increasing it for larger values.

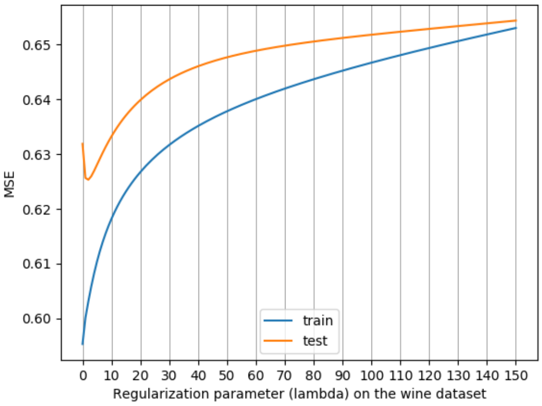
- How do you explain these variations? These variations would mostly depend on how different the test set samples are from the train set. If they are pretty similar, then the prior would only increase MSE. If they are different, then the prior would reduce MSE for the test set, but increase MSE on the train set.









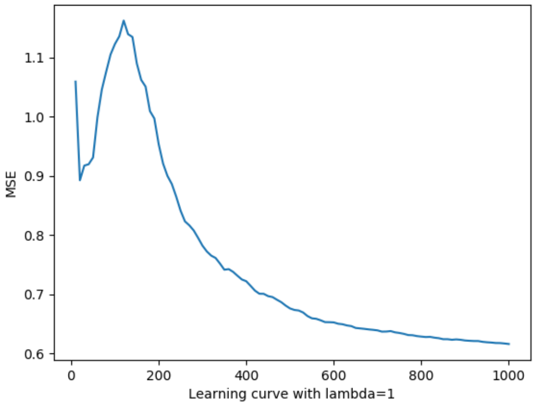


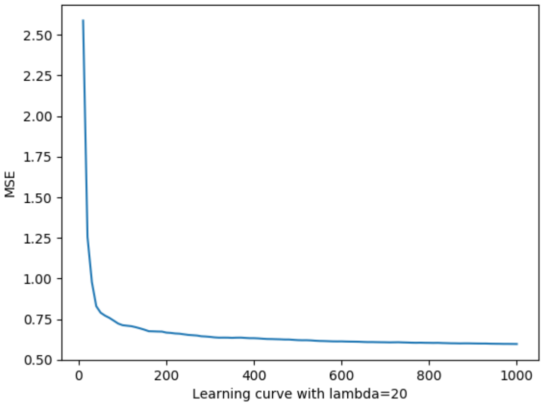
Task 2: Learning Curves

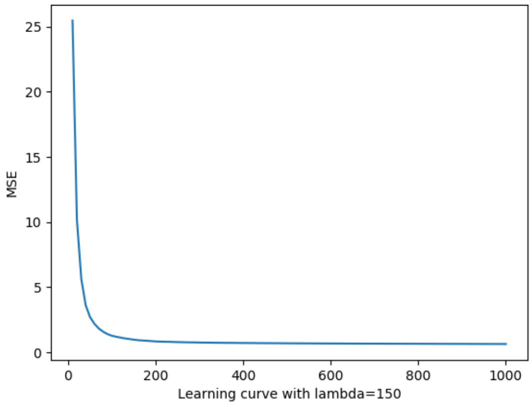
- What can you observe from the plots regarding the dependence of the error on λ and on the number of samples? We observe that the learning curves with the right λ gives the least error on train set sizes that are "just right". The condition for the train set size being "just right" is explained below.

  As the number of samples exceed a certain threshold required to capture the distribution correctly, extra samples do not give that much of a reduction on error. However, below that threshold, there is high error.

- Consider both the case of small training set sizes and large training set sizes. How do you explain these variations? Small train set sizes will not let us capture the distribution properly and hence we will have high error values. Large train set sizes will slow down our computation as the extra information does not contribute anything after the point when we have captured the distribution with enough train samples.







Task 3

Task 3.1: Parameter Selection using Cross-Validation

- 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?

  The choice of λ is different but the test set MSE is lower than that of part 1 as per the table below.

Task 3.2: Bayesian Model Selection

- 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?

  The choice of λ is different but the test set MSE is lower than that of part 1 as per the table below.

Dataset	Runtime (sec)	λ	Train set MSE	Test set MSE
100-10	0.8	17	6.308	6.249
100-100	7.2	15	0.687	0.722
1000-100	13.4	24	0.588	0.596
crime	8.9	150	0.343	0.392
wine	1.0	1	0.639	0.625

Table 1: Table of selected λ on different datasets using selection by cross-validation.

Dataset	Runtime (sec)	Test set MSE	λ
100-10	0.001	6.117	8.516
100-100	0.993	5.742	0.014
1000-100	0.012	0.608	0.5219
crime	0.017	0.389	58.446
wine	0.001	0.626	3.325

Table 2: Table of selected λ on different datasets using Bayesian Model Selection.

Task 3.3: Comparison

- How do the two model selection methods compare in terms of effective λ, test set MSE and run time?

  BME (Bayesian Model Estimation) is much faster than SCV (Selection by Cross-Validation), however SCV almost always gives a better Test set MSE than BME.

- Do the results suggest conditions where one method is preferable to the other?

  When we have enough computational power and time in our hands, we can opt for SCV as it has better performance. If we need a good estimate of parameters for cheap computational cost, we can opt for BME.