Task 1: Regularization

Q1. Provide results plots and discuss them.

Fig. 1

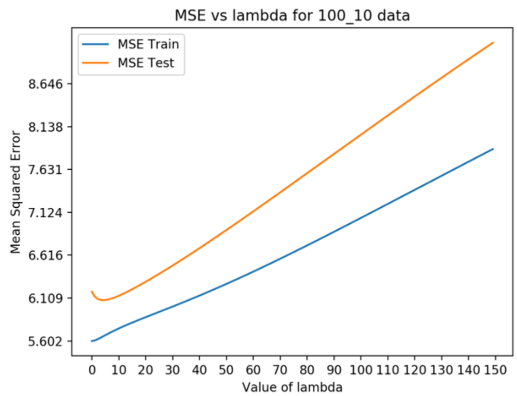


Fig. 1 shows the plot of test and training MSE for different values of λ. The data set has 100 examples with 10 features for each example. For smaller values of λ the gap between the training MSE and test MSE is small as the test MSE decrease on smaller values of λ, but as the value of λ increases the gap becomes larger because the rate of increase of test MSE is much higher than the rate of increase of training MSE, which suggest that increase in the value of λ, is making the model less accurate. It shows that at higher values of λ, the model is overfitting on the training data leading to increased test MSE. So, the optimal value of λ should be some initial values.

Fig. 2

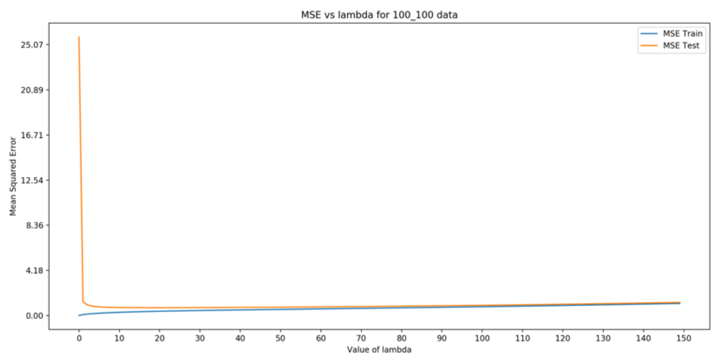


Fig. 2 shows the plot of test and training MSE for different values of λ. The data set has 100 examples with 100 features. For initial value of λ, the test error spikes to 25, then it decreases with increase of values of λ showing that λ is improving the test MSE. For mid-values of λ the test and train MSE remains consistent but both increases slowly, where test MSE shows more rate of increase with higher values of λ. It shows that the optimal value of λ should be some initial value.

Fig. 3

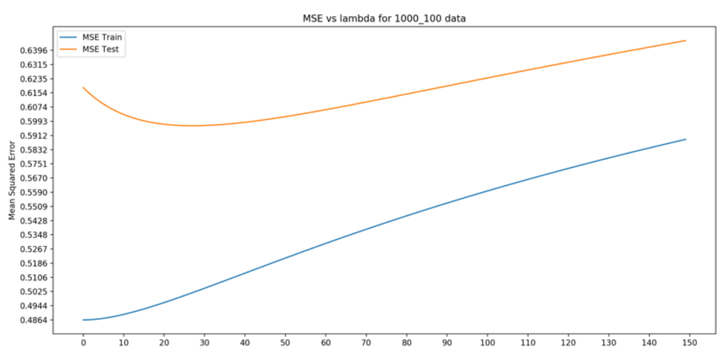


Fig. 3 shows the plot of test and training MSE for different values of λ. The data set has 1000 examples with 100 features. For initial values of λ, the train MSE and test MSE show significant difference. With slight increase in the value of λ, test MSE decrease to an optimal value and then starts to increase monotonically for higher values of λ. The training MSE also increase monotonically with values of λ. It shows the optimal values of λ should be some values in 20-30 range where test MSE is least.

Fig. 4

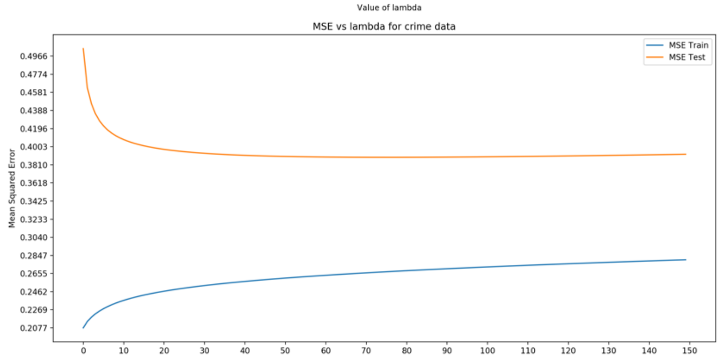


Fig. 4 shows the plot of test and training MSE for different values of λ for crime data set. For initial values of λ, the test MSE shows sharp decrease but with larger values of λ the rate of decrease of MSE slows but it continues decreasing, while the train MSE keeps increasing with the value of λ. The optimal values of λ should be some higher values where train and test MSE are less.

Fig. 5

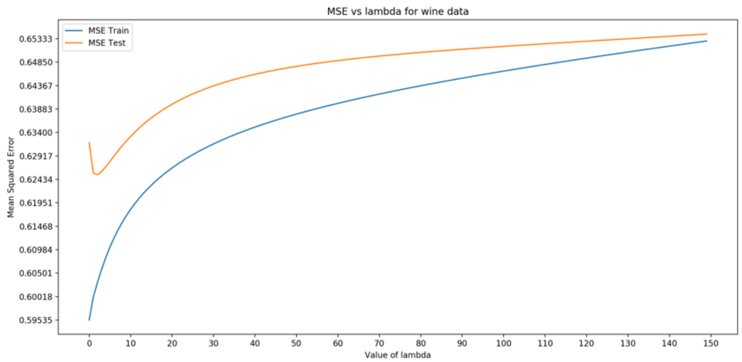


Fig. 5 shows the plot of test and training MSE for different values of λ for wine data set. For initial values of λ, the test MSE shows a sharp dip while the train MSE shows sharp increase. After the sharp din in test MSE it starts to increase with the increase in value of λ with decaying rate. At higher value of λ the training MSE and test MSE comes close with sharp increase in train MSE indicating the optimal value of λ is some smaller value.

Q2. Why can't the training set MSE be used to select λ.

Ans. The trend of training set MSE shows that it is monotonically increases for all the data sets, while the test MSE improves for certain values of λ to achieve some optimal MSE value and then starts to increase with higher values of λ. Therefore, for best performance of our model, the value of λ should be chosen on the basis of test MSE.

Q3. How does λ affect error on the test set?

Ans. The values of λ brings the test MSE towards some optimal MSE value by adding some penalty to its values. This improves the overall performance of the model.

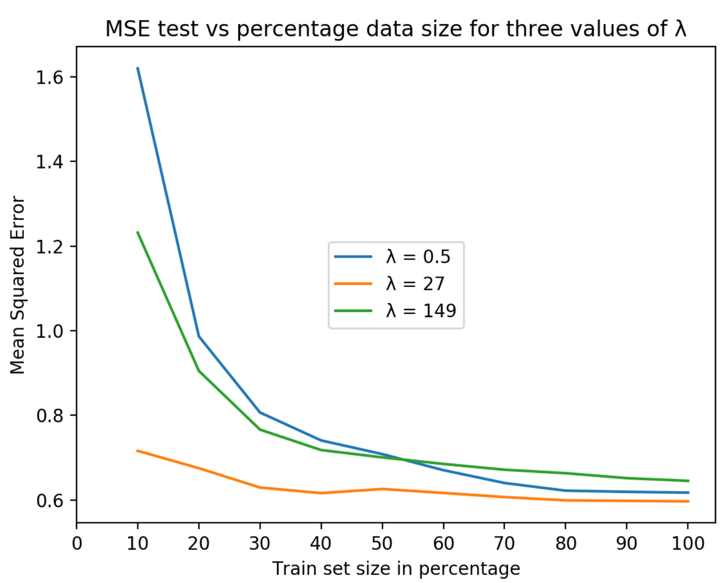
Q4. Does the value of λ differs for different data set? How do you explain this variation?

Ans. Yes, the optimal value of λ differs for different data set, the optimal value of λ depends on where the test MSE achieves it optimal value, which is the parameter to test accuracy of our models. Different data has different noises and patterns which affect the optimal value of test MSE, in turn affecting the optimal value of λ for the same data set.

Task 2: Learning Curves

Q1. Provide the plots and some numerical results and discuss them.

Fig. 6



Value of lambda	Min Test MSE	Max Test MSE
0.5	0.61731	1.619986
27	0.59674	0.7160712
149	0.64514	1.2317039

Table 1 showing min and max test MSE for different lambda values.

Fig 6 shows the test MSE error for three chosen values of λ (too small, just right and too large). As it can be observed from the plot as well as from the table 1. The for too small or too large values of λ, the model performance is not optimal and the minimum test MSE and maximum test MSE vary significantly. While for the just right value of λ, the minimum test MSE and Maximum test MSE are not significantly different, showing the best fit for the data set.

Q2. What can you observe from the plots regarding the dependence of the error on λ and on the number of samples? Consider both the cases of small training set sizes and large training set sizes, how do you explain these variations?

For too small or too large values of λ, the errors are high and model performance is not good on small training size of data. But as the training size is increased for them the test MSE decreases and comes quite close to test MSE of optimal λ. This shows that with larger training size, the model performance improves significantly. While for the optimal values of λ, the increase in training size data also improves the performance but for optimal λ, the performance at smaller data set are also far better. For all data size, the performance of model with optimal λ is best.

Task 3.1: Model selection using cross validation

Q1. Apply the method to 5 datasets and report the values of λ selected, associated MSE and the run time.

Name of data set	Test MSE	Lambda value	Run time (sec)
100_10	6.2144	15	0.904
100_100	0.7203	18	1.507
1000_100	0.597	23	1.789
Crime	0.3923	149	1.576
Wine	0.6259	3	1.197

Table 2 shows test MSE, λ and run time for model selection using cross validation.

Q2. 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?

Name of data set	Test MSE	values of lambda
100_10	6.0847	5
100_100	0.7202	19
1000_100	0.5967	28
Crime	0.389	76
Wine	0.6253	3

Table 3 test MSE and lambda for part 1

As compared with the test MSE values of part 1, the test MSE values of this part if very close to each other. The test MSE values are almost same for the both process but the optimal value of λ is not exactly same as of part 1. The values of λ in part 1 and this part are around similar range, like, both have comparable values of lambda for 100-10, 100-100, 1000-100 and wine data sets. But for wine data set the value of lambda is almost halved but follows the same overall trend for λ as well.

Task 3.2: Bayesian Model Selection

Q1. Apply this to 5 datasets and report the values of alpha, beta, the effective λ, the associated MSE and the run time.

Name of data set	Alpha	Beta	Effective Lambda	Test MSE	Run time (sec)	No of iteration to converge
100_10	0.8821	0.1652	5.3407	6.088	0.033	6
100_100	5.1546	3.1543	1.6342	1.0635	0.291	13
1000_100	10.2858	1.8603	5.5291	0.6083	0.104	4
Crime	425.6448	3.2504	130.9502	0.3911	0.271	12
Wine	6.1639	1.6098	3.829	0.6267	0.084	19

Table 4. shows the values of alpha, beta, λ, test MSE, run time and no of iteration to converge for Bayesian model selection.

Q2. 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?

Ans: As shown in table 3 and table 4, there are different interpretation for different types of data sets, for small data set with less number of features (like, 100-10, crime and wine) the values of test MSE for part 1 and Bayesian method comes same, also for data set with larger number of examples and more features, similar things happens. The test MSE comes almost close in both cases. But Bayesian method do not perform well with less examples and having a greater number of features, like, 100-100 data set. The test MSE for this case comes much larger than the part 1 suggesting bad performance for smaller data sets with a greater number of features. The values of lambda comes smaller for all the cases by Bayesian method. Specially, where Bayesian method is performing bad, the value of lambda is much smaller like in the case of 100-100 data set.

Task 3.3: Comparison

Q1. How do the two model selection methods compare in terms of effective λ, test set MSE and run time?

	Cross validation method			Bayesian method		
Name of data set	Test MSE	Lambda value	Run time (sec)	Test MSE	Lambda value	Run time (sec)
100_10	6.2144	15	0.958	6.088	5.3407	0.033
100_100	0.7203	18	1.528	1.0635	1.6342	0.291
1000_100	0.597	23	1.839	0.6083	5.5291	0.104
Crime	0.3923	149	1.595	0.3911	130.9502	0.271
Wine	0.6259	3	1.213	0.6267	3.829	0.084

Table 5. shows the values of λ, test MSE, run time for Bayesian model selection and cross validation method.

Ans: For all data sets, Bayesian model run time is much smaller than the cross validation run time showing that this model is faster in execution. For most of the cases the test MSE is similar but Bayesian model fails for smaller data set with larger number of features and gives larger test MSE and much smaller value of lambda. For other data sets the values are comparable for both the cases, and test MSE is close enough for both.

Q2. Do the results suggest conditions where one method is preferable to the other? Try and thing about the results obtained and discuss these questions even if you do not see an obvious trend.

Ans: Bayesian method do not perform well with less examples and having a greater number of features, like, 100-100 data set. The test MSE for this case comes much larger than the part 1 suggesting bad performance for smaller data sets with a greater number of features. This suggests that cross validation method should be preferred over Bayesian method for model selection in such cases. For other cases, the Bayesian method is much faster in running time and produces comparable or better result, so it should be preferred.