Assignment2:

Task 1:

1.1 Crime dataset:

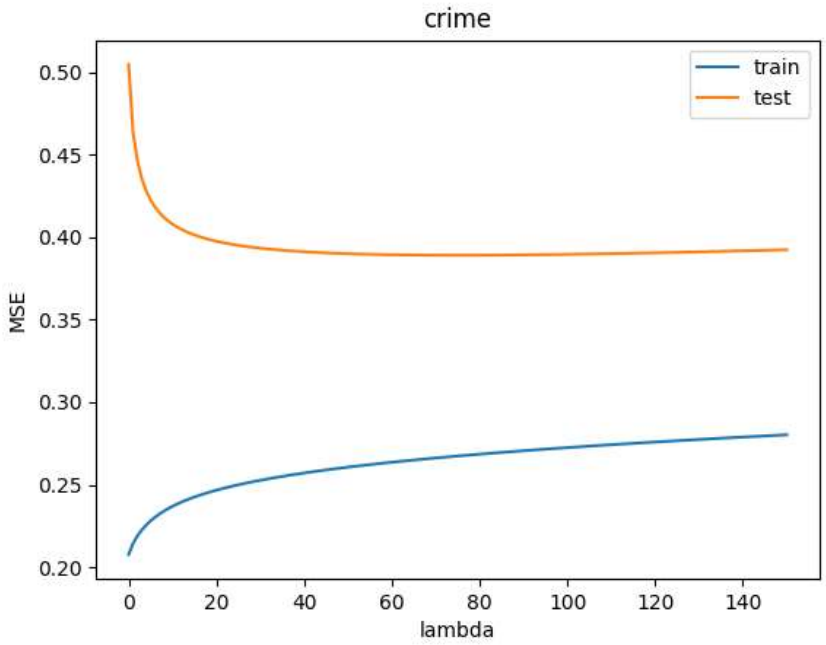

Figure 1:MSE based on the lambda in crime dataset

In this dataset, the best MSE (lowest one) on the test dataset is 0.3890233877134439 and the lambda in the range 0 to 150 that gives this result is 75.

best lambda: 75

best MSE: 0.3890233877134439

number of training samples: 298

number of features: 100

By adding lambda, we reduce the high variance and overfitting. But actually we have some difference between train errors and test errors.

The gap in errors between training and test suggests a high variance problem in which the algorithm has overfit the training set. ***Reducing the feature set*** will ameliorate the overfitting and help with the variance problem. Also, ***adding more training data*** will increase the complexity of the training set and help with the variance problem.

Now

1.2. wine dataset:

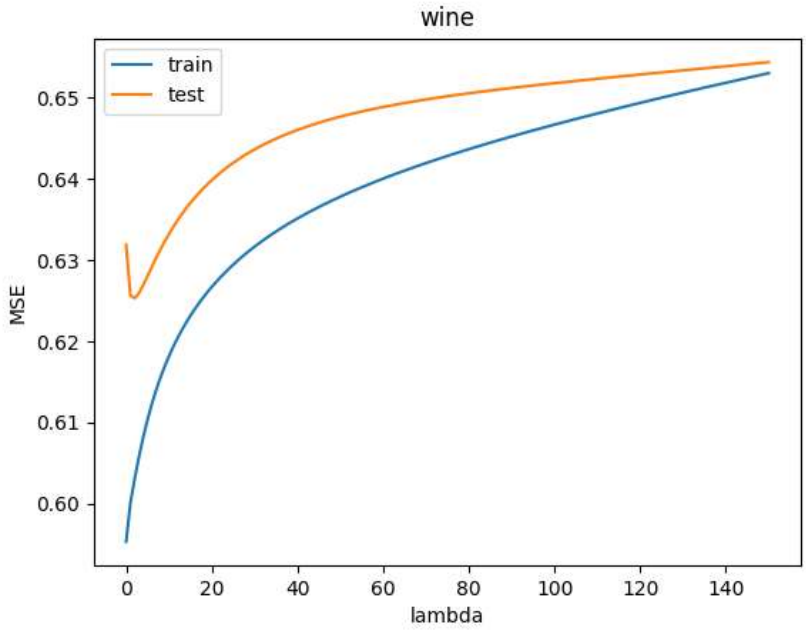

Figure 2: MSE based on the lambda in wine dataset

In this dataset, the best MSE (lowest one) on the test dataset is 0.6253088423047256 and the lambda in the range 0 to 150 that gives this result is 2.

best lambda: 2

best MSE: 0.6253088423047256

number of training samples: 342

number of features: 11

1.3. artlarge dataset

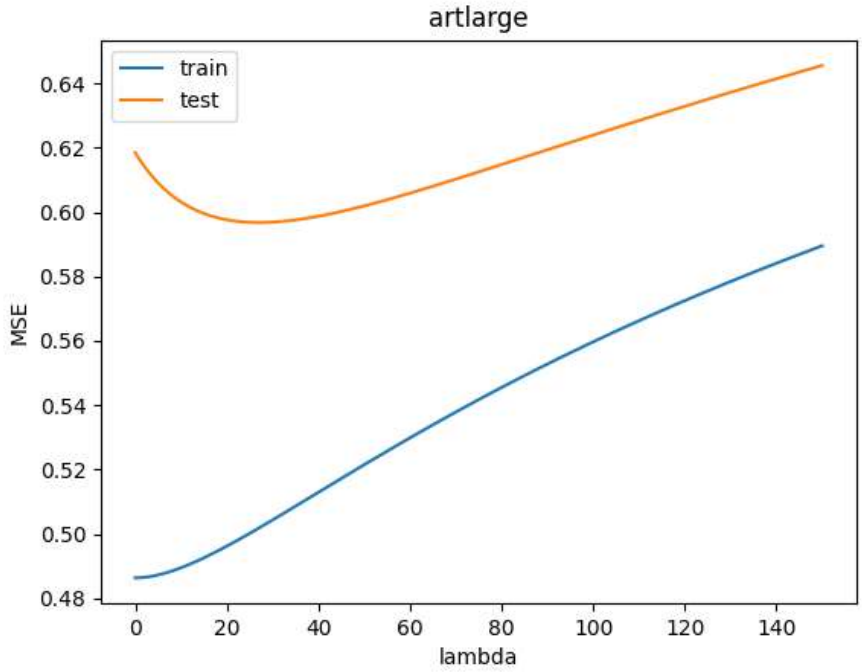

Figure 3: MSE based on the lambda in artlarge dataset

In this dataset, the best MSE (lowest one) on the test dataset is 0.5967438457326988 and the lambda in the range 0 to 150 that gives this result is 27.

best lambda: 27

best MSE: 0.5967438457326988

MSE of the true function: 0.533

These 2 MSE's are close so we are on the right track

number of training samples: 1000

number of features: 100

1.4. artsmall dataset:

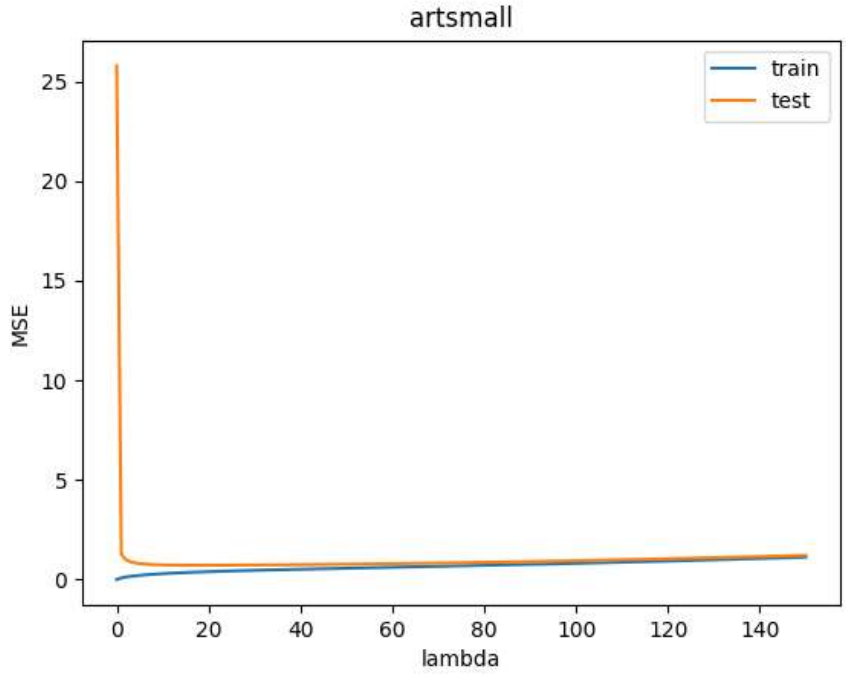

Figure 4: MSE based on the lambda in artsmall dataset

In this dataset, the best MSE (lowest one) on the test dataset is 0.7202788056527144 and the lambda in the range 0 to 150 that gives this result is 18.

best lambda: 18

best MSE: 0.7202788056527144

MSE of the true function: 0.557

These 2 MSE's are close so we are on the right track.

number of training samples: 100

number of features: 100

The learning algorithm finds parameters to minimize training set error, so the performance should be better on the training set than the test set

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

The λ that gives the best MSE for training datasets is 0 because λ add some noise so the model won't be the best model for the dataset to get the minimum MSE.

We should not use training error to choose the regularization parameter, as we can always improve training error by using less regularization (a smaller value of lambda). But too small of a value will not generalize well on the test set.

Q: How does λ affect error on the test set?

The λ avoid overfitting by adding some noise (overfitting has 'High-variance').

It adds some bias therefore prevent overfitting. So, we get better MSE for test sets by choosing the proper lambda.

Q: Does this differ for different datasets? How do you explain these variations?

There is an optimum point in each dataset for lambda. Before that we have a high variance and after that we have a high bias.

Actually, choosing a good lambda always help to get better results and the shapes are conceptually the same for all datasets. But the best value of the lambda is different from one dataset to another dataset.

The best value of lambda depends on the datasets and number of features and the scale and values of the features and also it depends on the number of samples in the dataset.

Training good, test bad:

The gap in errors between training and test suggests a high variance problem in which the algorithm has overfit the training set. Reducing the feature set will ameliorate the overfitting and help with the variance problem. Also, adding more training data will increase the complexity of the training set and help with the variance problem. Adding a good lambda can help us with that too.

Training and test both bad:

The poor performance on both the training and test sets suggests a high bias problem. Increasing the regularization parameter will allow the hypothesis to fit the data worse, decreasing both training and test set performance.

Lambda:

We are adding a small bias to the over-fitting cost function (which tends to zero) to prevent it from overfitting. By adding that small amount of bias, we get a significant drop in the variance.

Task2:

Q. 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?

Usually, we don't have a test dataset results so we can not use the test datasets to find the best lambda so the cross validation is a good way to test our results on the training dataset.

The MSE of test sets are better for the previous task because we check all lambdas and see which one gives us best MSE for test datasets. But in this task, we just checked our answer on the training set and we get best lambdas based on that.

But the best MSE and best lambda in this task are close enough to task 1 which means we are on the right track and we can use this method too.

2.1. crime dataset:

best lambda: 150

best MSE: 0.39233899203438116

2.2 wine dataset:

best lambda: 3

best MSE: 0.6259362266964954

2.3 artlarge dataset:

best lambda: 7

best MSE: 0.6063621582528399

2.4: artsmall dataset:

best lambda: 4

best MSE: 0.8338953861335517

Task 3:

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?

Usually, we don't have a test dataset results so we cannot use the test datasets to find the best lambda so the cross validation is a good way to test our results on the training dataset.

The MSE of test sets are better for the task1 because we check all lambdas and see which one gives us best MSE for test datasets. But in this task, we just checked our answer on the training set and we get best lambdas based on that.

But the best MSE and best lambda in this task are close enough to task 1 which means we are on the right track and we can use this method too.

3.1. crime dataset

alpha: 521.3760222544001

beta: 0.3594058678702361

MSE: 0.4797055914325741

lambda: 1450.6608513210017

3.2. wine dataset

alpha: 15.780621230138351

beta: 0.6400225603618894

MSE: 0.6418350494214086

lambda: 24.656351521758044

3.3. artlarge

alpha: 9.633199060479049

beta: 0.5383036208149019

MSE: 0.5982479972685675

lambda: 17.895475133338305

3.4 artsmall

alpha: 4.556692912828721

beta: 0.4569811071491861

MSE: 0.7360342238045963

lambda: 9.971293870888939

Task 4:

	Task 1	Task 2	Task 3
crime test best MSE	0.3890233877134439	0.39233899203438116	0.3911023069944971
crime best Lambda	75	150	130.95038735098254
crime runtime	—	3.1192076206207275	0.39977121353149414
crime bast alpha	—	—	425.64534067367697
crime best Beta	—	—	3.250432085648071
wine test best MSE	0.6253088423047256	0.6259362266964954	0.6267461680483954
wine best Lambda	2	3	3.828946654298746
wine runtime	—	0.8165304660797119	0.15690302848815918
wine bast alpha	—	—	6.163874068692461
wine best Beta	—	—	1.6098093353618024
artlarge test best MSE	0.5967438457326988	0.6063621582528399	0.6083085920181194
artlarge best Lambda	27	7	5.529076503114884
artlarge runtime	—	5.713715553283691	0.3647909164428711
artlarge bast alpha	—	—	10.285792777281662
artlarge best Beta	—	—	1.860309361154076
artsmall test best MSE	0.7202788056527144	0.8338953861335517	1.0635170829308733
artsmall best Lambda	18	4	1.6341288080412573
artsmall runtime	—	2.4875717163085938	0.2748417854309082
artsmall bast alpha	—	—	5.154599496905394
artsmall best Beta	—	—	3.1543409990329567

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

Effective lambda: The best lambdas I got from Task 2 are nearer to Task 1 rather than Task 3 which means Task 2 is better and more accurate in terms of finding better Lambda than evidence approximation

MSE: The best MSE's I got from Task 2 are better than Task 3 which means Task 2 is better and more accurate in terms of finding better MSE than evidence approximation.

So, in general Task 2 gives us a better answer rather than task 3.

Run time: run time of the evidence approximation is less because it does not need to calculate a lot and it's less expensive.

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

When the dataset is too large and just the result is important not the model parameter it is better to use evidence approximation since the other one is too expensive to calculate and take more time. (if the difference between two runtimes are significant)

We do not have a test dataset usually so the second and the third method should be used.

If we want a better MSE, (lower MSE), we have to use task2.