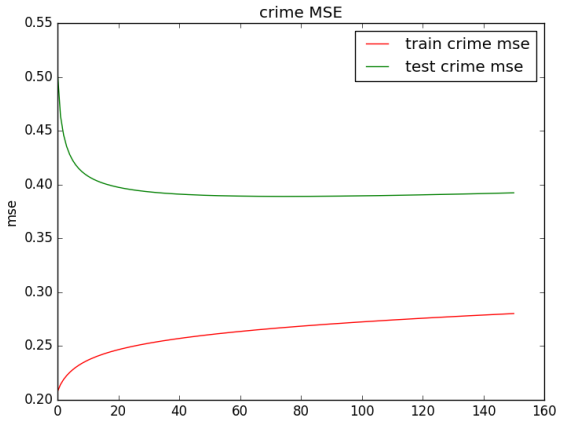
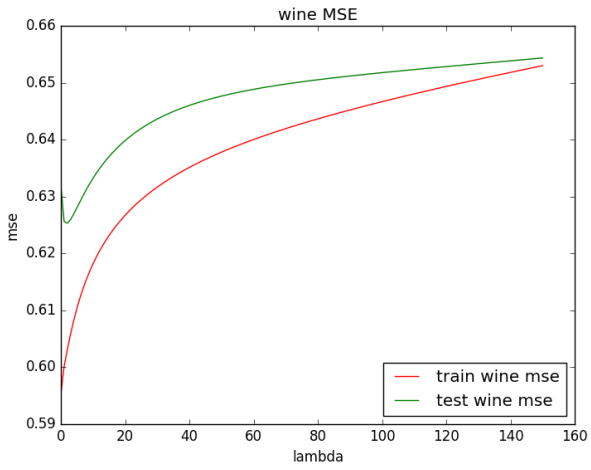
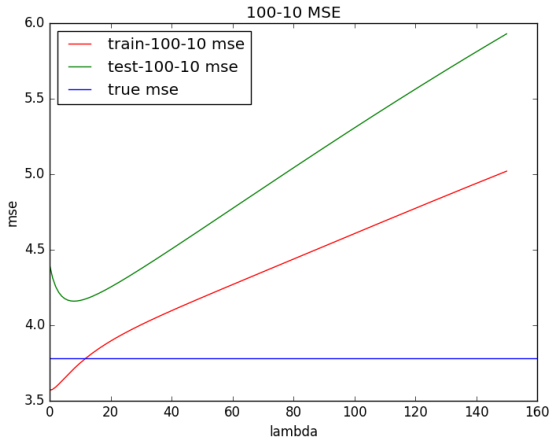
COMP136 PP2 report

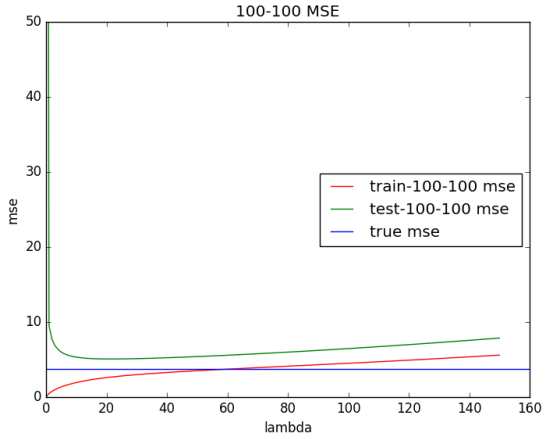
By Beibei Du
10/27/2016

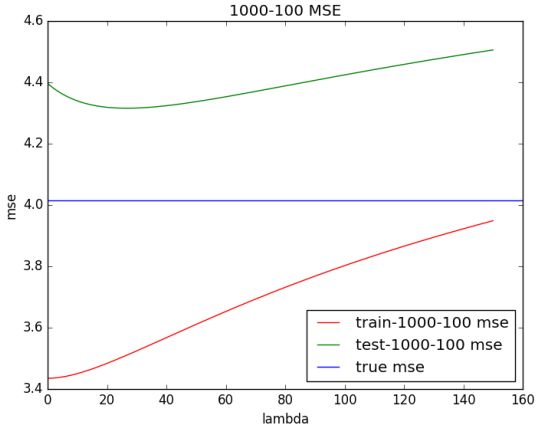
Task 1: Regularization

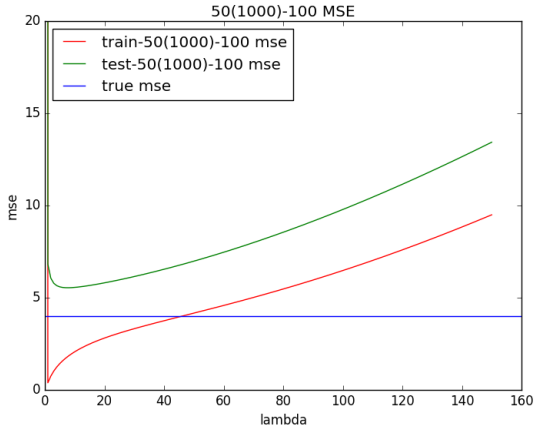


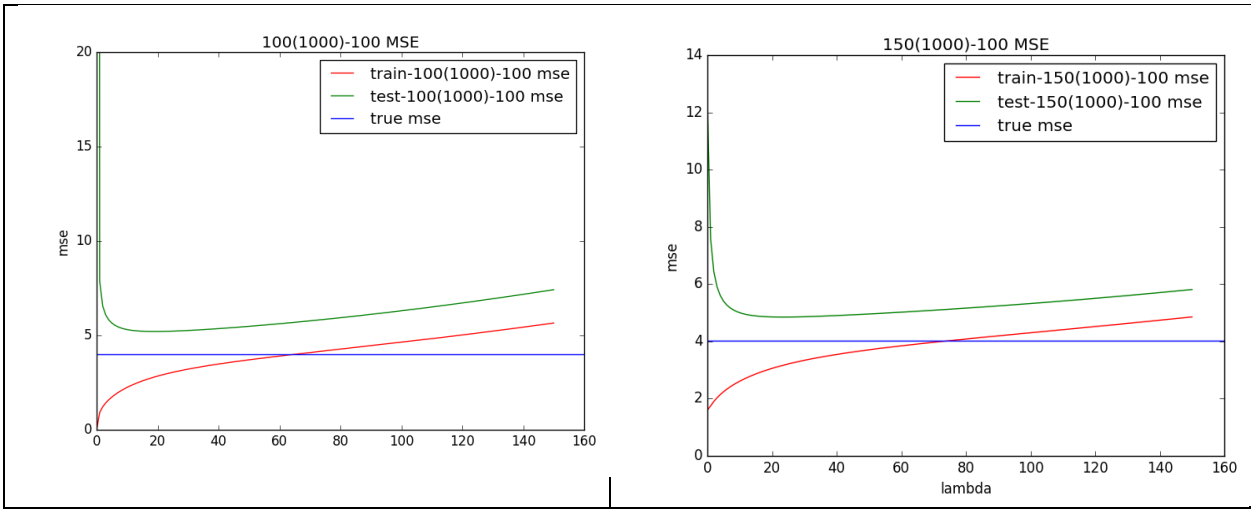












1. **Why can't the training set be used to select lambda?**

   Answer:

   According to figures above, we can find that training data MSE increasing with the increasing value of lambda for most of data sets, that means when lambda = 0, it has the minimum MSE, so we cannot use training set to select lambda.

2. **How does lambda affect error on the test set?**

   Answer:

   According to figure above, we can find that for all 8 datasets, testing MSE first decrease, then increase with the increasing value of lambda, for artificial dataset, it is first approaching the true MSE, then go away, which means that at some point(lambda), we can find minimum testing MSE with an optimal lambda. For larger lambda, our model may have overfit.

   For these datasets, the optimal lambda and its MSE shows as below:

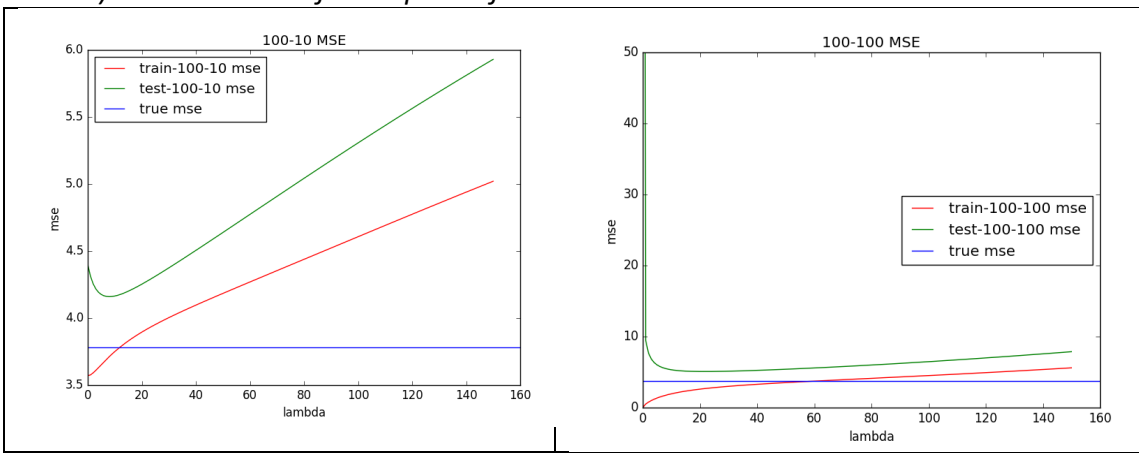
Dataset	Optimal lambda	Test MSE
Crime	75	0.389023387713
Wine	2	0.625308842305
100-10	8	4.15967850948
100-100	22	5.07829980059
1000-100	27	4.31557063032
50(1000)-100	8	5.54090222919
100(1000)-100	19	5.20591195733
150(1000)-100	23	4.84894305335

3. **How does the choice of the optimal lambda vary with the number of features and number of examples? The number of features is fixed, the number of examples is fixed. How do you explain these variations?**

Answer:

Following above table including optimal lambda and its Test MSE with this lambda:

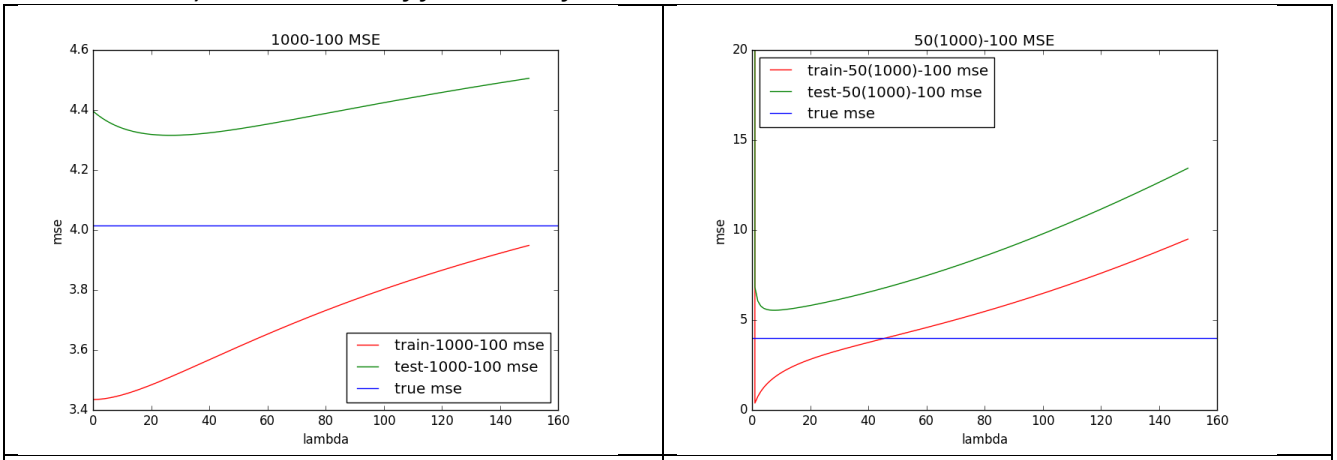
a) The number of examples is fixed



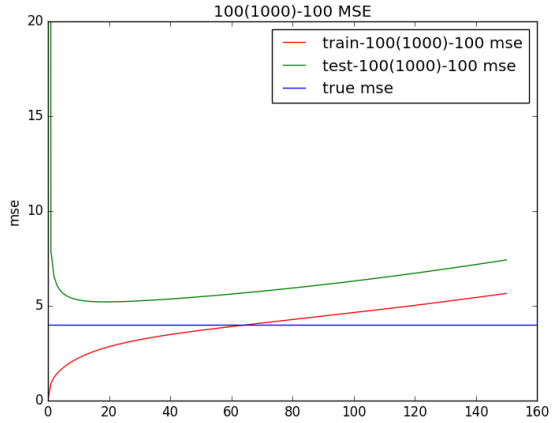
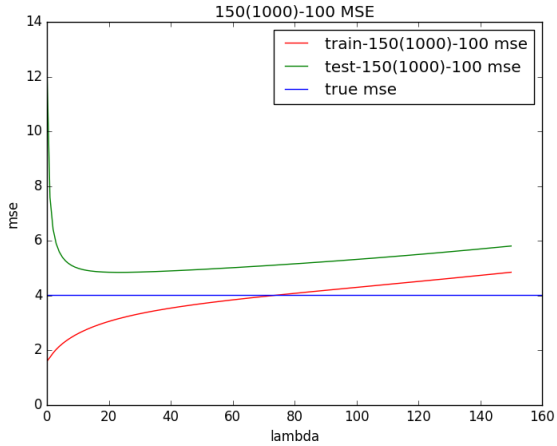
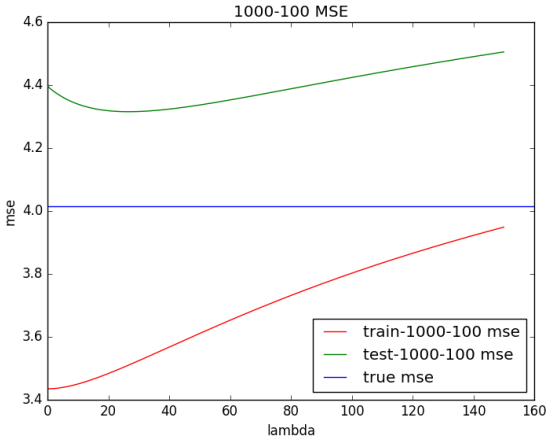
Data Set	Optimal lambda	Test MSE
100-10	8	4.15967850948
100-100	22	5.07829980059

Data set 100-10 and 100-100 have same number of examples but different number of features, we can find that 100-100 optimal lambda is greater than 100-10 optimal lambda, which means that optimal lambda is increasing with the increase of number of features with same examples.

b) The number of features is fixed



Optimal lambda: 27 , min MSE: 4.31557063032	Optimal lambda: 8, min MSE: 5.54090222919

<table>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>Optimal lambda: 19, Min MSE:5.20591195733</td><td>Optimal lambda: 23, min MSE: 4.84894305335</td></tr>
<tr><td> 
</td><td></td></tr>
<tr><td>Optimal lambda: 22, min MSE: 5.07829980059</td><td></td></tr>
</table>

Data set 50(1000)-100, 100(1000)-100, 150(1000)-100, 1000-100 have same number of features but different number of examples, according to figure and their optimal lambda comparisons, we can find that optimal lambda is increasing with the increase of number of examples. This is also can be verified comparing data set 100-100 and 1000-100, 1000-1000 optimal lambda is also greater than 100-100 optimal lambda. They all shows with fixed number of features, optimal lambda will increase if number of examples increase.

c) *How do you explain these variations?*

Number of examples fixed: comparing dataset 100-10 and 100-100.

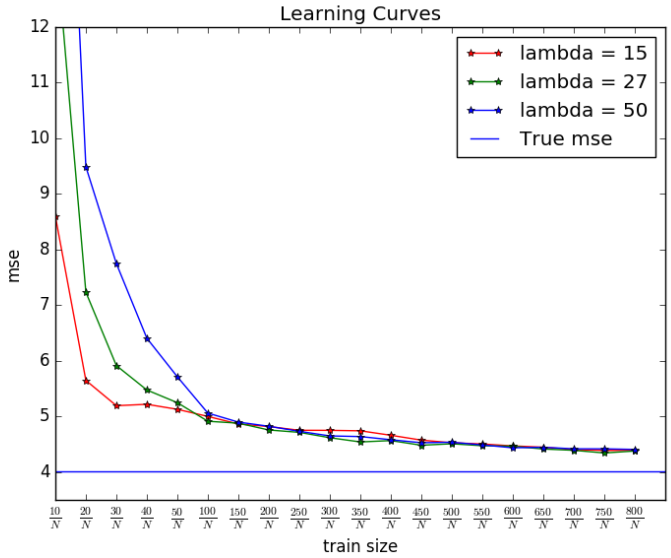
100-10 with few features is dominated by lambda, so with smaller lambda, it behaves better, but for dataset 100-100 with more features, it can absorb the effects of a larger lambda.

Number of features fixed: comparing dataset 50(1000)-100, 100(1000)-100, 150(1000)-100.if dataset has fewer examples, which means that it was dominated by lambda, so with smaller lambda, its performance is better. If

dataset has more examples, the case can absorb larger lambda, also larger lambda has the benefit of normalizing the data to a better performance.

Task2: Learning Curves

1. **What can you observe from the plots regarding the dependence on lambda and the number of examples? Consider both the case of small training set sizes and large training set sizes. How do you explain these variations?**



Answer:

For dataset 1000-100, the optimal lambda is 27, so I select represent lambda 15, 27 and 50, and plot figure as above.

We can find test MSE will decrease with the increase of train size, independent of lambdas, and also the test MSE is above true MSE for all lambdas, even optimal lambda.

a) Small train set size

According to figure, we can find that larger MSE with larger lambda, which means that, when train size is smaller, lambda have bigger effect on test MSE, the ratio of lambda is bigger, the test MSE is also bigger. This is because when the train set size is smaller, this case is dominated by lambda, smaller lambda can give better performance.

b) Large train set size

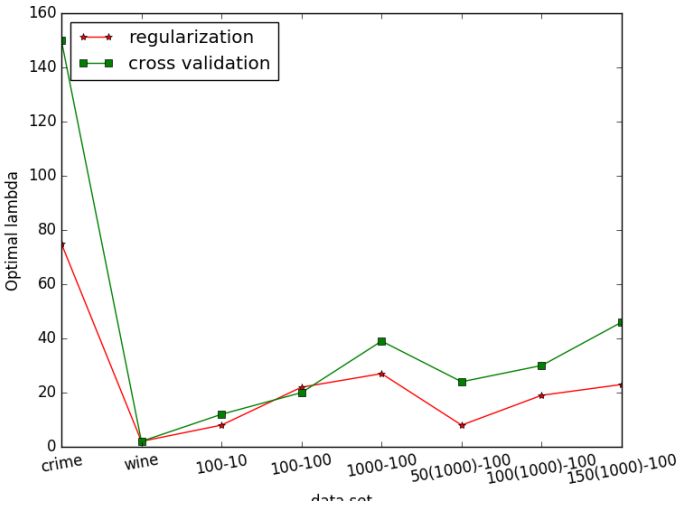
According to figure, we can find that with large train set size, the ratio of lambda becomes smaller, test MSE of three lambdas become almost the same. Which also means that with large train set size, test MSE is independent of lambda. This is because with larger data set size, it can absorb the effect of larger lambda, the lambda matters less.

Task3: Cross Validation

1. **How do the results compare to the best test-set results from part 1 both in terms choice of lambda and test set MSE? What is the run time cost of this scheme?**

	Cross Validation		Part 1	
Dataset	Optimal lambda	Test MSE	Optimal lambda	Test MSE
Crime	150	0.392338	75	0.3890233
Wine	2	0.6253088	2	0.6253088
100-10	12	4.175709	8	4.1596785
100-100	20	5.080888	22	5.0782998
1000-100	39	4.322722	27	4.3155706
50(1000)-100	24	5.934465	8	5.5409022
100(1000)-100	30	5.259982	19	5.2059119
150(1000)-100	46	4.9341926	23	4.84894305

a) Compare best test-set results in terms choice of lambda



Cross validation and regularization give us different optimal lambda.
Figure above is to compare optimal lambda between part1 regularization and cross validation, we can find that for most of 8 datasets, cross validation optimal lambda is greater than or equal to regularization optimal lambda.

b) Compare best test-set results in terms choice of test set MSE

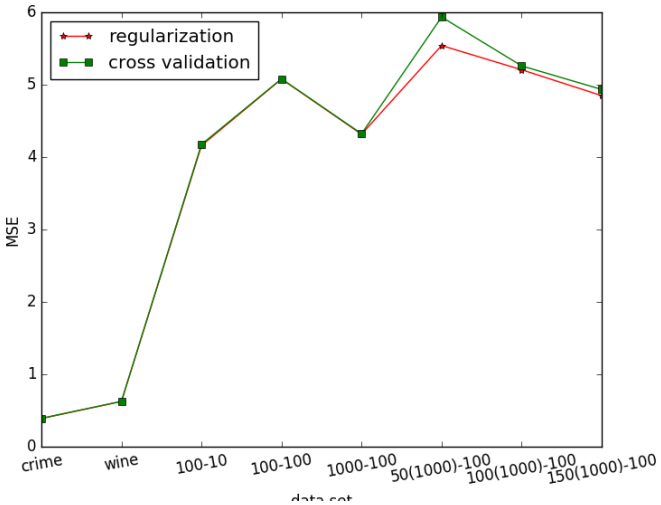


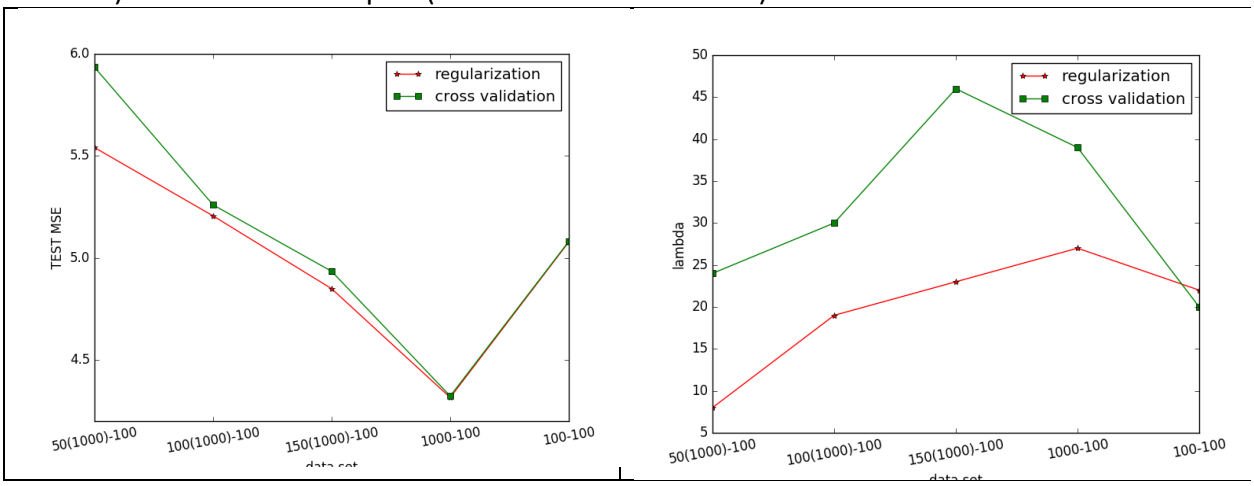
Figure above is to compare best test-set results in terms choice of test set MSE. We can find that cross validation test MSE is greater than or equal to regularization best test MSE. When the number of examples is smaller, cross validation performs not better than regularization. Back to cross validation methods, the calculation method is same for both of these two methods, but instead of using train set once to pick parameter w, cross validations compute ten times of that and pick average value as w, if number of examples smaller, it cannot pick a better w than regularization.

c) Run time cost of this scheme
Lambda : range(151) m, Folder: range(10) n
Total time cost is about O(1500), generalized as O(m*n)
It's larger than regularization which is linear.

Dataset: [crime, wine, 100-10, 100-100, 1000-100, 50(1000)-100, 100(1000)-100, 150(1000)-100]
run time: [1.850564, 0.214724, 0.128529, 1.296833, 3.708794, 4.776038, 6.029359, 7.403829]

2. How does the quality depend on the number of examples and features?

a) Number of examples (fixed number of features)



From figure above, data set 50(1000)-100, 100(1000)-100, 150(1000)-100, 1000-100 have same number of features.

Cross Validation Method: if number of examples increase, test MSE will decrease.
if number of examples increase, optimal lambda will first increase, then decrease.

Part1 regularization: if number of examples increase, test MSE will decrease.
If number of examples increase, optimal lambda will increase.

The MSE difference between Cross Validation method and part1 regularization, if number of examples increase, the difference will become smaller.

b) Number of features (fixed number of examples)
Comparing data 100-10 and 100-100, which have same number of examples but different number of features.

	Cross Validation		Part 1	
Dataset	Optimal lambda	Test MSE	Optimal lambda	Test MSE
100-10	12	4.175709159671	8	4.15967850948
100-100	20	5.080888817918	22	5.07829980059

According to data as above table, we can find that optimal lambda will increase if number of examples increase for both of cross validation method and regularization, same as test MSE will also increase.

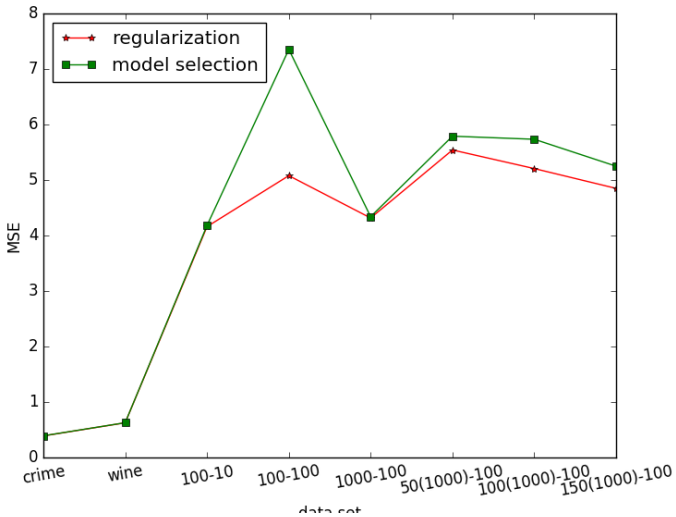
Difference between cross validation and part1, if number of features increase, difference will also become smaller.

Task4: Bayesian Model Selection

1. **How do the results compare to the best test-set results from part 1 both in terms choice of lambda and test set MSE? What is the run time cost of this scheme?**

	Model Selection		Part 1	
Dataset		Test MSE	Optimal lambda	Test MSE
Crime		0.391102208062	75	0.389023387713
Wine		0.626744824930	2	0.625308842305
100-10		4.18013303120	8	4.15967850948
100-100		7.35253692507	22	5.07829980059
1000-100		4.33834995092	27	4.31557063032
50(1000)-100		5.78957529375	8	5.54090222919
100(1000)-100		5.73393073453	19	5.20591195733
150(1000)-100		5.24899661548	23	4.84894305335

Type equation here.



a) Since we don't have optimal lambda for Bayesian Model selection, so just compare best test results in choice of test MSE, and the figure shows as above.

According to figure, we can find that for all of these 8 datasets, Bayesian Model selection test MSE is greater than or almost equal to part1 test MSE.

b) Run time cost of this scheme

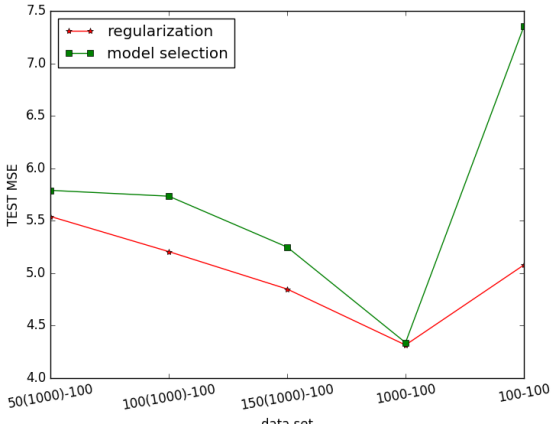
Bayesian model selection is different than regularization, here we give alpha and beta one initialized value, using convergence when alpha and beta won't change, that we find the parameter mN to calculate MSE. So the time cost is the time that how many times that we find the mN, of course it also depends on the size of data set, but here we ignore that. I just use one counter to see the number of convergence.

Data set	crime	wine	100-10	100-100	1000-100	50(1000)-100	100(1000)-100	150(1000)-100
counter	11	13	3	13	3	8	6	2

run time: [0.151428, 0.034536, 0.007779, 0.105999, 0.117914, 0.154577, 0.183432, 0.19713]

2. How does the quality depend on the number of examples and features?

a) Number of examples (fixed number of features)



From figure above, data set 50(1000)-100, 100(1000)-100, 150(1000)-100, 1000-100 have same number of features.

Bayesian Model Selection: if number of examples increase, test MSE will decrease.

Part1 regularization: if number of examples increase, test MSE will also decrease.

Difference MSE: if number of examples increase, difference between them will decrease.

b) Number of features (fixed number of examples)

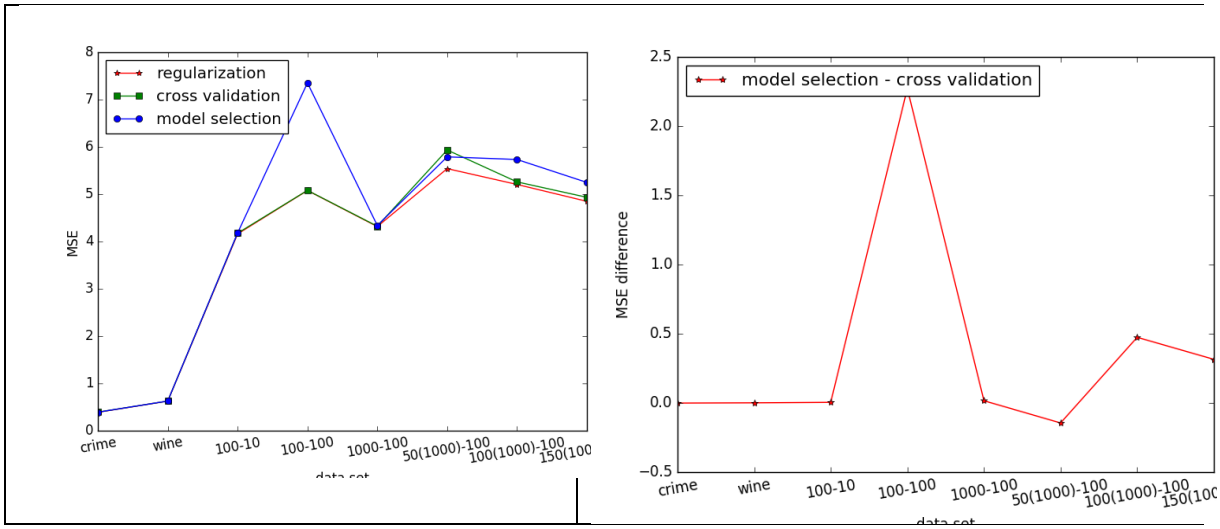
	Model Selection		Part 1	
Dataset		Test MSE	Optimal lambda	Test MSE
100-10		4.18013303120	8	4.15967850948
100-100		7.35253692507	22	5.07829980059

Compare 100-10 and 100-100 data sets, with fixed number of examples, if number of features increase, we can find that for both Bayesian Model selection and Part1, test MSE will also increase. Difference between them will increase if number of features will increase.

Task5: Comparison

1. **How do the two model selection methods compare in terms of the test set MSE and in terms of run time?**

	Model Selection		Cross Validation	
Dataset		Test MSE	Optimal lambda	Test MSE
Crime		0.391102208062	150	0.392338
Wine		0.626744824930	2	0.625309
100-10		4.18013303120	12	4.175709
100-100		7.35253692507	20	5.080888
1000-100		4.33834995092	39	4.322722
50(1000)-100		5.78957529375	24	5.934465
100(1000)-100		5.73393073453	30	5.259982
150(1000)-100		5.24899661548	46	4.9341926



a) Compare in terms of the test set MSE
Comparing the MSE of these data sets, shown as figure above, Model selection MSE is greater than or equal to cross validation MSE. And we can find that for data set 100-100, the difference is biggest (2.0 < difference < 2.5). But when dataset is 50(1000)-100, model selection has smaller MSE than cross validation, which means that Bayesian model selection has better performance if the dataset is smaller.

b) Compare in terms of run time

Regularization is linear in time complexity;
Cross validation time complexity is O(m*n) m is lambda range, n is number of folders.
Bayesian run time is much faster than cross validation according to below table.

Dataset	Cross Validation run time (seconds)	Bayesian model selection (seconds)
Crime	1.850564	0.151428
Wine	0.214724	0.034536
100-10	0.128529	0.007779
100-100	1.296833	0.105999
1000-100	3.708794	0.117914
50(1000)-100	4.776038	0.154577
100(1000)-100	6.029359	0.183432
150(1000)-100	7.403829	0.19713

2. **What are the important factors affecting performance for each method? Given these factors, what general conclusions can you make about deciding which model selection method to use?**

Answer:

Cross Validation: lambda, number of examples and features

Bayesian Model selection: number of examples and features

It's easier to do the calculations for cross-validation, comparing to compute posterior probabilities using Bayesian model selection.
Considering number of examples, we can find that if when number of examples is smaller, cross validation performs better, if the purpose is to get better model, then cross validation is a good choice. If number of examples is larger, comparing data set 1000-100, Bayesian model selection can also have better performance, since cross validation time cost is high, if don't have enough time, Bayesian model selection is a better choice.
Considering number of features, 100-10 and 100-100, with larger number of features, cross validation performs much better than Bayesian model selection, so if small number of features, and also want to save time, we can use Bayesian model selection, but if larger number of features, it's better use cross validation.