Daniel Campos November 7th , 2013

RCS: Campod2 RIN:660996361

Machine Learning CS 4100

I worked on these problems with Zoe Konrad☺

Problem Set 9

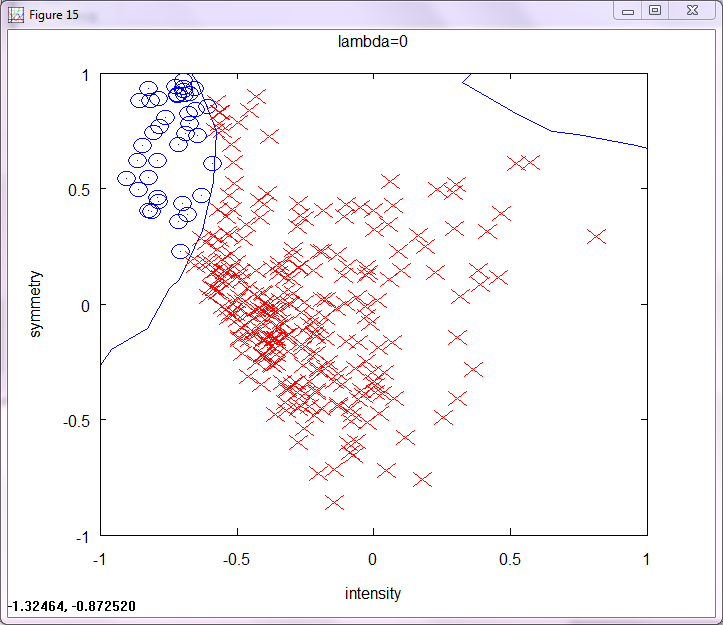
1. (100) 8th order Polynomial Feature Transform.

Use the 8th order polynomial feature transform to compute Z. What are the dimensions of Z?

The dimensions for the resulting matrix Z are 300(the amount of data sets I am testing by 45(the amount of terms in an 8th order expansion)

1. (100) Overﬁtting.

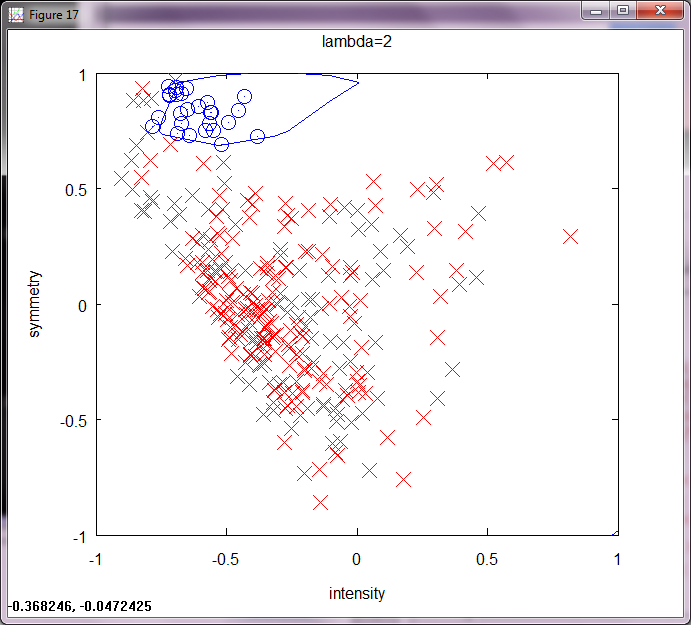
Give a plot of the decision boundary for the resulting weights without any regularization (λ = 0). Do you think there is overﬁtting or underﬁtting?



There is over fitting to a decent amount. All the points are correctly classified.

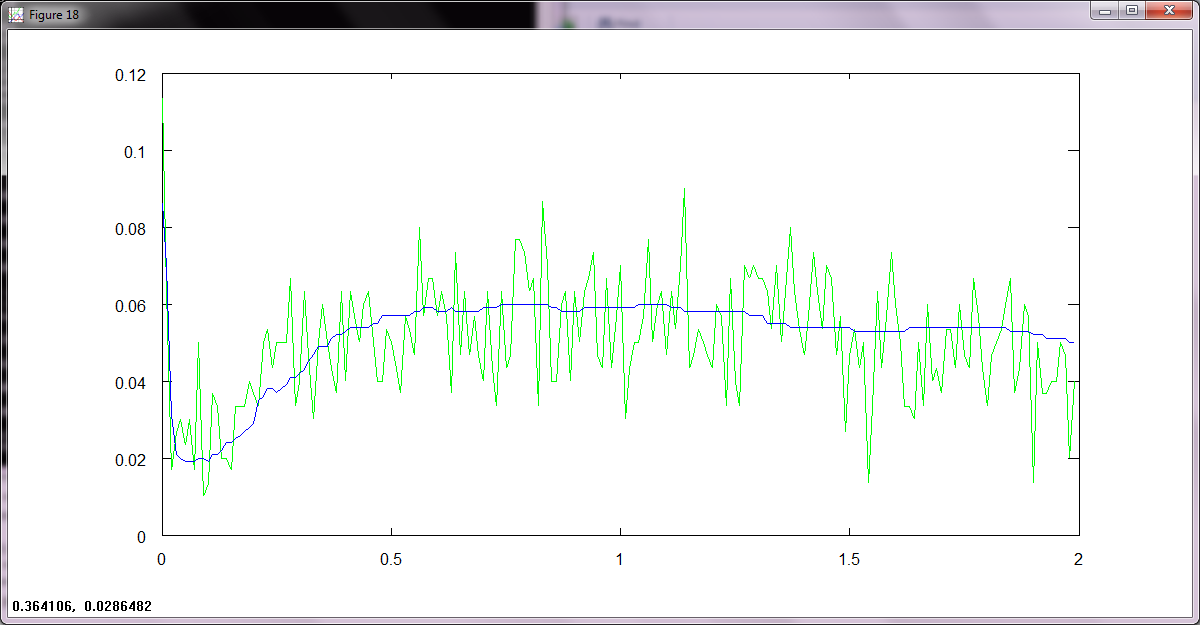
1. (100) Regularization

Give a plot of the decision boundary for the resulting weights with λ = 2. Do you think there is overﬁtting or underﬁtting?



There is under fitting in Lambda=2 since there are many points that are not classified correctly.

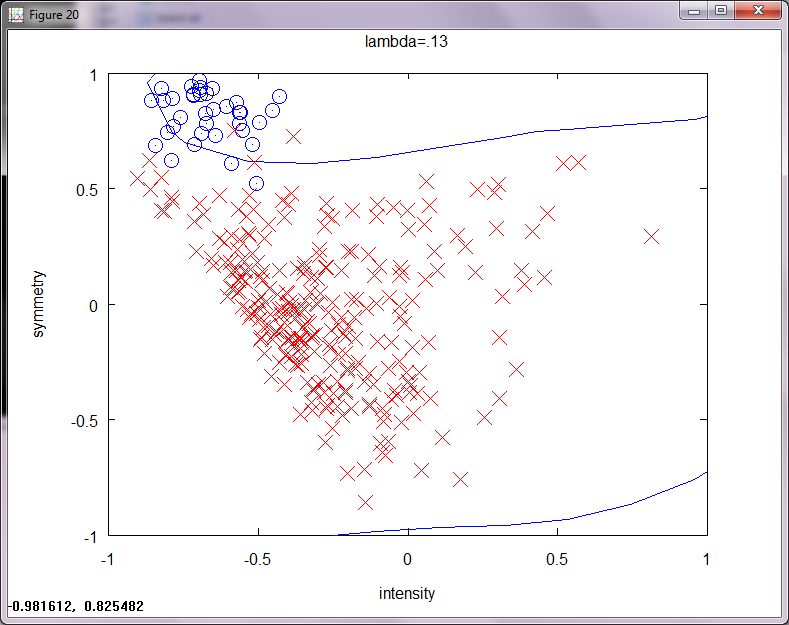
1. (200) Cross Validation.



Ecv follows much of what Etest does but because it is random spikes up and down but besides that follows the eTest but has a lower average value despite its spikes.

1. (100) Pick λ.

I picked a λ∗ of .13 bellow is its decision boundary plot



1. (100) Estimate Eout.
   1. Based on theorem 2.12

 🡪 √((8/1000)ln((4(2001))/.05)=.3005366 so we are 30% confident that etest will be within eout by 5%

with a bound of 5%

The estimate for eout using etest for my final hypothesis is

Ein=.016714 from matlab

etest=.022000 from matlab

so based on the bound it is correct. Thus we can apply this bound to the rest of the points and it follows what we get. When I did all the data points I got

eOut=.019754 so withn the bound and close to eTestπ

1. (100) Is Ecv biased?

Yes it is biased. It is biased because we were selecting our λ∗ based on the results that the test set returned which means we data snooped. We are allowed to do anything we want with the training data before there is snooping. We went beyond that.

1. (200) Data snooping.

It is not an unbiased estimate because we did data snoop. We used our test set to select which λ∗ we were going to use and thus have data snooped. As you said online “You can do anything you want with the training data. Here is a very simple prescription that you can use and it will never let you down:  
  
Take your test data and lock it up in a password protected encrypted file to which only your client has the password. (Note: you can be your own client.)  
  
Now do *whatever* you want with the training data to obtain your final hypothesis [http://book.caltech.edu/vblatex/img/b2f5ff47436671b6e533d8dc3614845d-1.gif](javascript:;). Give it to the client. When the client asks you what performance to expect with [http://book.caltech.edu/vblatex/img/b2f5ff47436671b6e533d8dc3614845d-1.gif](javascript:;), you ask her to open the test data file and run your [http://book.caltech.edu/vblatex/img/b2f5ff47436671b6e533d8dc3614845d-1.gif](javascript:;) on that file. The result on the test data is the performance to expect. The client is now stuck with that [http://book.caltech.edu/vblatex/img/b2f5ff47436671b6e533d8dc3614845d-1.gif](javascript:;) and that test performance. You are not allowed to change [http://book.caltech.edu/vblatex/img/b2f5ff47436671b6e533d8dc3614845d-1.gif](javascript:;) any more.” We took the test set and selected our final g, which was λ∗, based on the test data that we should not have had access to. Bad.