Jiahuan He
ID: 912490740

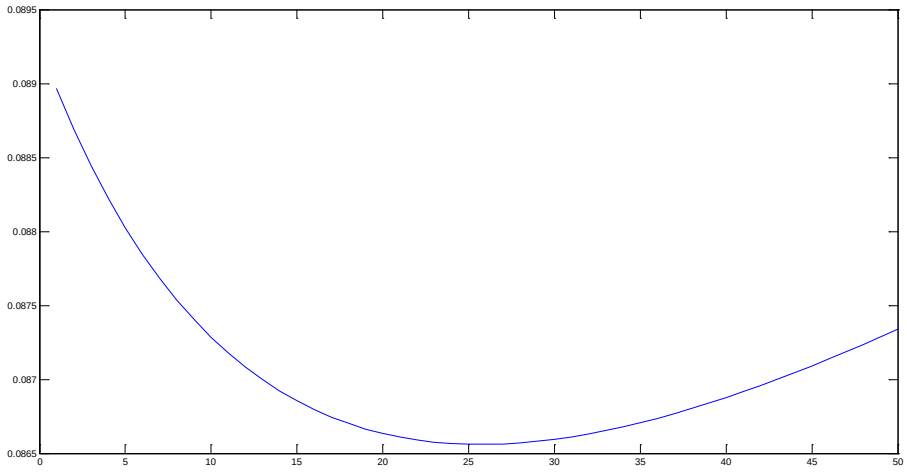
##The homework data is uploaded into Matlab workspace in two parts – object "data" includes GrowthRate and all the genes while "textdata" contains ID, Strain, Medium, Stress and GenePerturbed.

#1.(see "rr.m", "hw3q1q2.m". "rr.m" is the self-defined function that output the coefficients estimated by ridge regress with preset lambda)

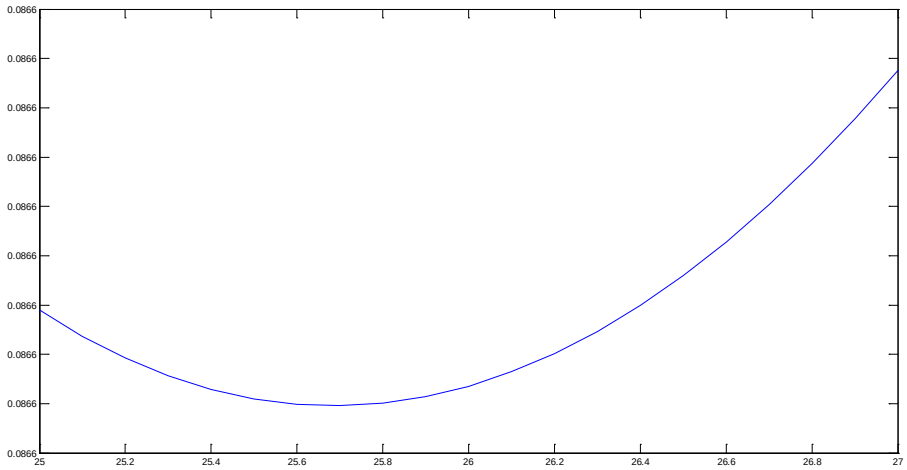
In this question, we are going to apply ridge regression. Compare to the ordinary least squares' RSS: $\mathrm{RSS} = \sum_{i=1}^{m}(y^{(i)} - w^T x^{(i)})^2$, ridge regression adds up a term of $\lambda \sum_{j=0}^{n} w_j^2 \quad (\lambda > 0)$ into RSS to penalize the complexity of the model. λ is so called constrained parameter. Ridge regression provides a mechanism to "eliminate" unimportant X variables as well as a remedy for multicolinearity. There is explicit solution for parameters of ridge regression:

$$\widehat{w}(\lambda) = (X^T X - \lambda I)^{-1} X^T Y$$

There is no universally best way to find what should λ be equal to. An intuitive method is to analyze the ridge trace plot. λ is considered to be good when the curves of $\widehat{w}_j$s become stable. In this case, however, we have more than 4000 X variables, thus it is impossible to apply this method. Alternately, we will just use 10-fold cross-validation to find the λ with the least MSE. Note that although $\widehat{w}(\lambda)$ is a biased estimation of w, the MSE could be smaller than that of the regular linear regression.
First, we try $\lambda = 1, 2, \ldots 50$



We can see that a λ in somewhere between 20 and 30 will produce the least MSE. Actually, MSE decreases until 26. So, similarly, we try $\lambda = 25, 25.1, \ldots 27$, and we have:



Thus, we decide to choose **25.7** as constrained parameter λ. Its MSE with 10-fold cross-validation is 0.0866. 61 of the coefficients are exactly 0. These are corresponding to those genes with all 0 values. The absolute value of 391 genes' coefficients are smaller than 1×10^{-4}, 2244 of them are smaller than 1×10^{-3}.

#2(see "hw3q3.m")

The prediction is 0.5430.

#3(see "svmmulti.m", "getlevel.m", "hw3q3.m". "svmmulti.m" is the self-defined function that output multiple hyperplanes. "getlevel.m" can output a list of all the categories of a variable.)

In building SVM, we will only use the genes with absolute value of coefficient bigger than 1×10^{-3}, which is indicated in #1. Since we are building a multiple classifier, we will apply one-verse-all method. We should also note that by doing cross-validation, it is possible that the testing set include categories which has no observation in training set. This might affect the accuracy of the SVM.
Here is the accuracy for Strain, Medium, Stress and GenePerturbed:

Labels	Accuracy
Strain	86.14%
Medium	87.84%
Stress	81.98%
GenePerturbed	90.54%