Homework 3: Regularization

Matej Kalc

Additional tests

Additional tests I implemented for Ridge and Lasso regression are:

1. Test if Ridge and Lasso return almost the same output as a linear function (y(x) = 10 + 2x and y(x) = 10 - 2x)
2. Test if Ridge and Lasso return almost the same output if $\lambda = 0$
3. Test if Ridge and Lasso return different output for $\lambda = 1$ and $\lambda = 100$
4. Test if Ridge and Lasso return different output for different input

Data preprocessing

All the features have been standardized using the mean and standard deviation of the train data set.

Results

The optimal λ for the superconductor application was selected by minimizing the RMSE of the test data set using the Powell minimize. Results are shown in Table 1. Since generally the test set is not known an appropriate λ should be chosen using only the train data set. Using CV (Cross Validation) I can obtain some info on how the model will behave on different lambdas. For our task, I choose 10-fold CV and tried to minimize the λ using Powell minimization. Using λ achieved with 10-fold CV gives worse results then the optimal λ (see Table 2) on the test data set. On the train set, it performed better. As I increase the number of folds, RMSE lowers. The λ provided with CV is a better guess then the optimal λ in Table 1, since in general it would perform better. I have to note that the whole data set is really small (only 300 instances). A bigger data with CV would provide a λ, which generally will perform better.

λ	99.64
RMSE on train set	18.52
RMSE on test set	14.76

Table 1. Results of the optimal λ on test data set

λ	4.66
RMSE on train set	14.39
RMSE on test set	18.37

Table 2. Results of the optimal λ with 10-fold CV on train data set