Data Analysis and Knowledge Discovery, Assignment 2

Niko Hellgren, 505174

December 20, 2016

1 *k*-NN prediction

The model used was programmed in Python with little to no use of external libraries. The data was z-score standardized, pre-calculated into a sample distance matrix, which was then used to leave-one-out cross-validate the data. The CV-data was used to tune the value of k to minimize the mean-square error (MSE) in the data. Since the classification is into discrete categories, the distance from the data point to the range that rounds to the correct value was considered in the calculation.

The results of the tuning process of k is presented in Figure 1. Based on the results, the value $k = 12$ gives the lowest error, albeit with little margin to the other low values. The cross-validation results of 12-NN are presented in Figures 2 and 3.

Despite the relatively low value of MSE, the results are not too good. The uneven distribution of samples in different qualities causes most samples to be predicted to be of quality 6, which makes up 44.9% of the whole data set, with the range 5–7 making up 77.9%. The low correlation between the features and the quality (as was seen in assignment 1) causes the evenly distributed 6-samples in the train set to pull most test samples to that class.

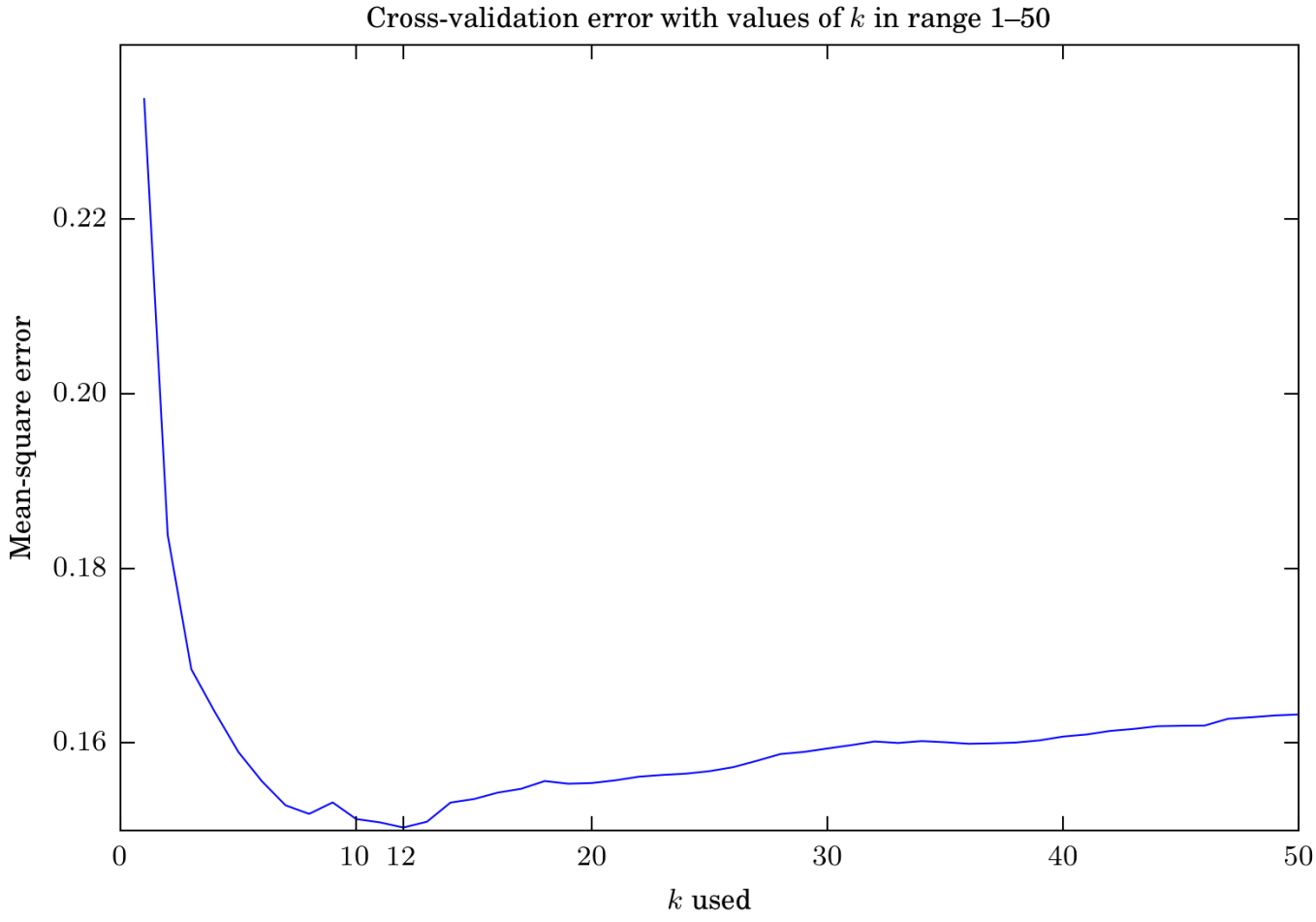

Figure 1: Errors of k-NN cross-validation

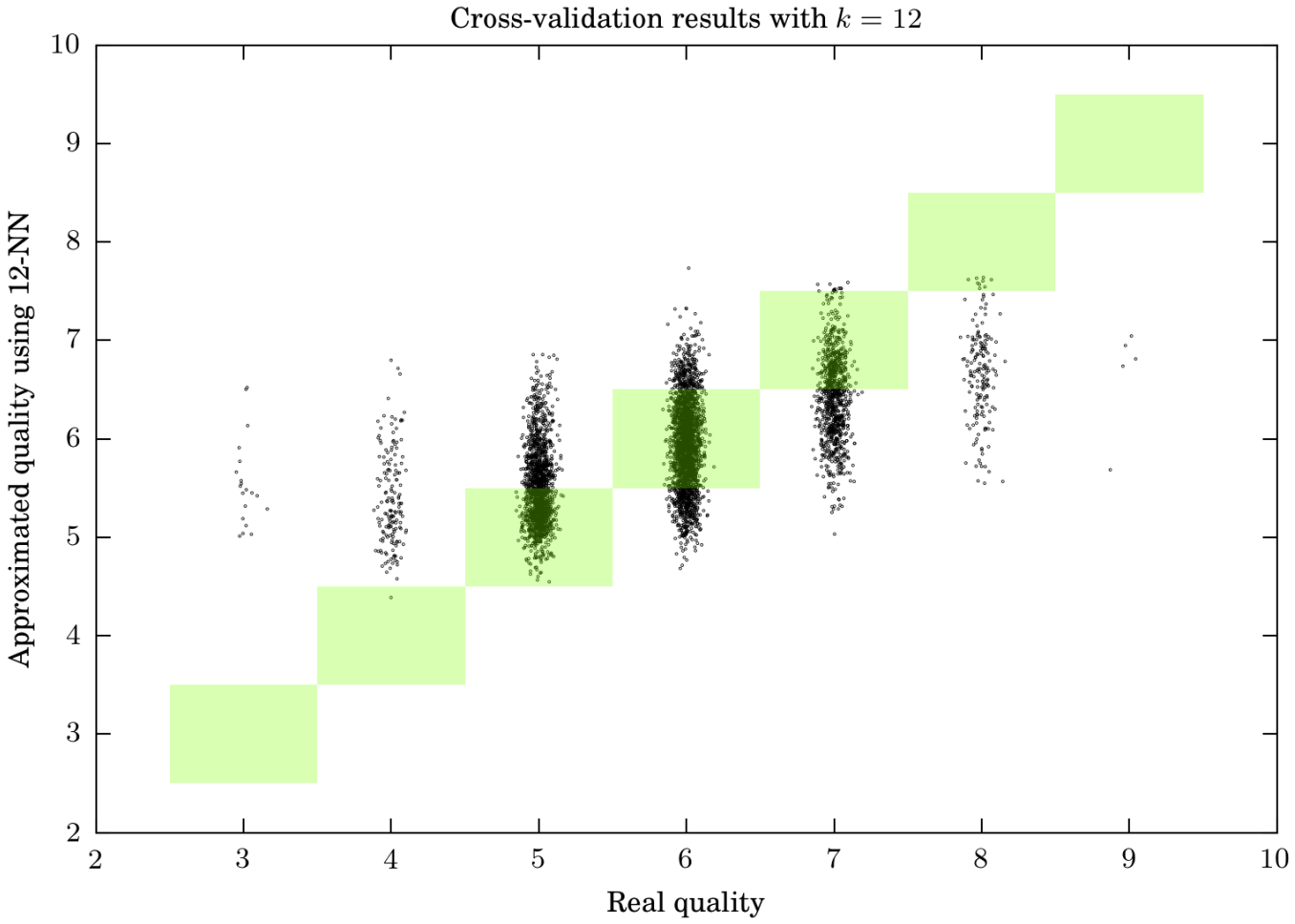

Figure 2: Results of 12-NN cross-validation as a scatter plot. The green boxes represent areas where the prediction is correct. All points have jitter following N(0,0.05) to avoid overlapping.

Figure 3: Confusion matrix of 12-NN cross-validation

2 Linear regression

Regression model was built using the examples presented in the lectures for Ridge regression. The same standardized data was used and leave-one-out cross-validated to determine a good value for the hyperparameter λ.

The results of the tuning process of λ are presented in Figures 4 and 5. Due to the random ordering of the samples, the values of λ with the smallest error varied between runs, but was often the smallest with values around 40. The cross-validation results of Ridge regression with $\lambda = 40$ are presented in Figures 6 and 7.

The MSE values are larger than with k-NN, and it can be seen as worse predictions. The samples in class 6 were very well predicted, but this is partially caused by the model predicting most of the test samples as quality 6.

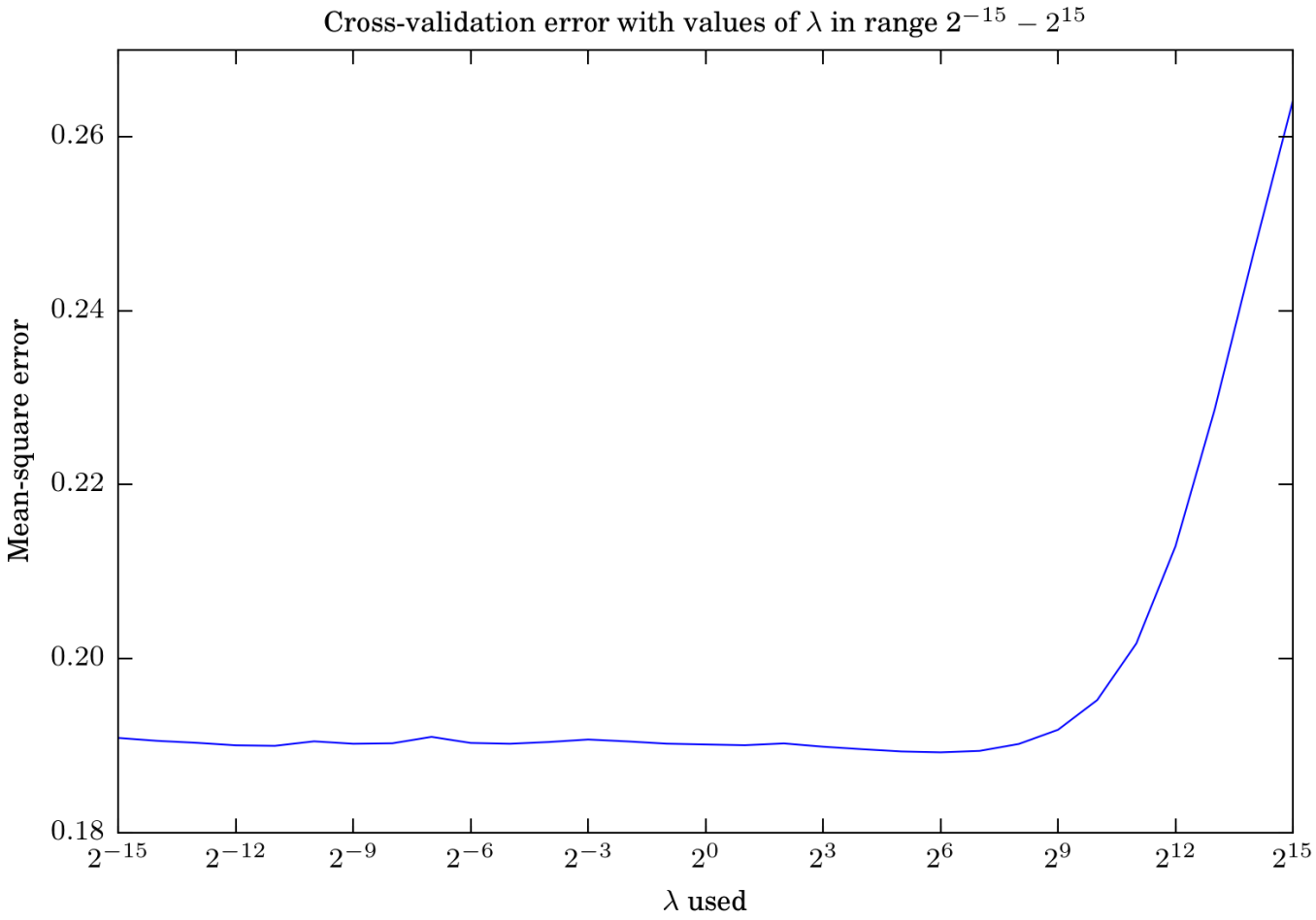

Figure 4: Errors of Ridge regression cross-validation

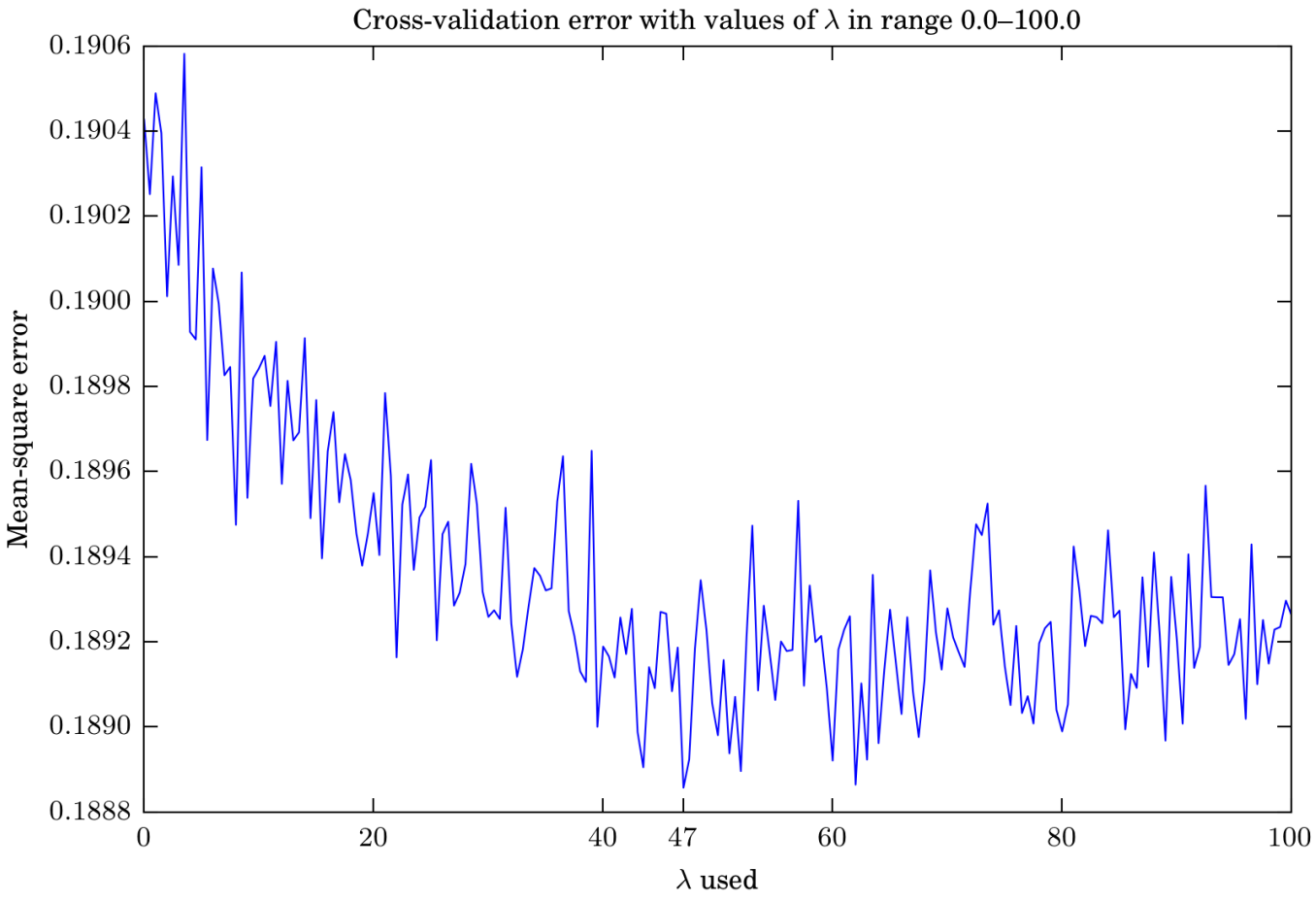

Figure 5: Errors of Ridge regression cross-validation in a smaller range

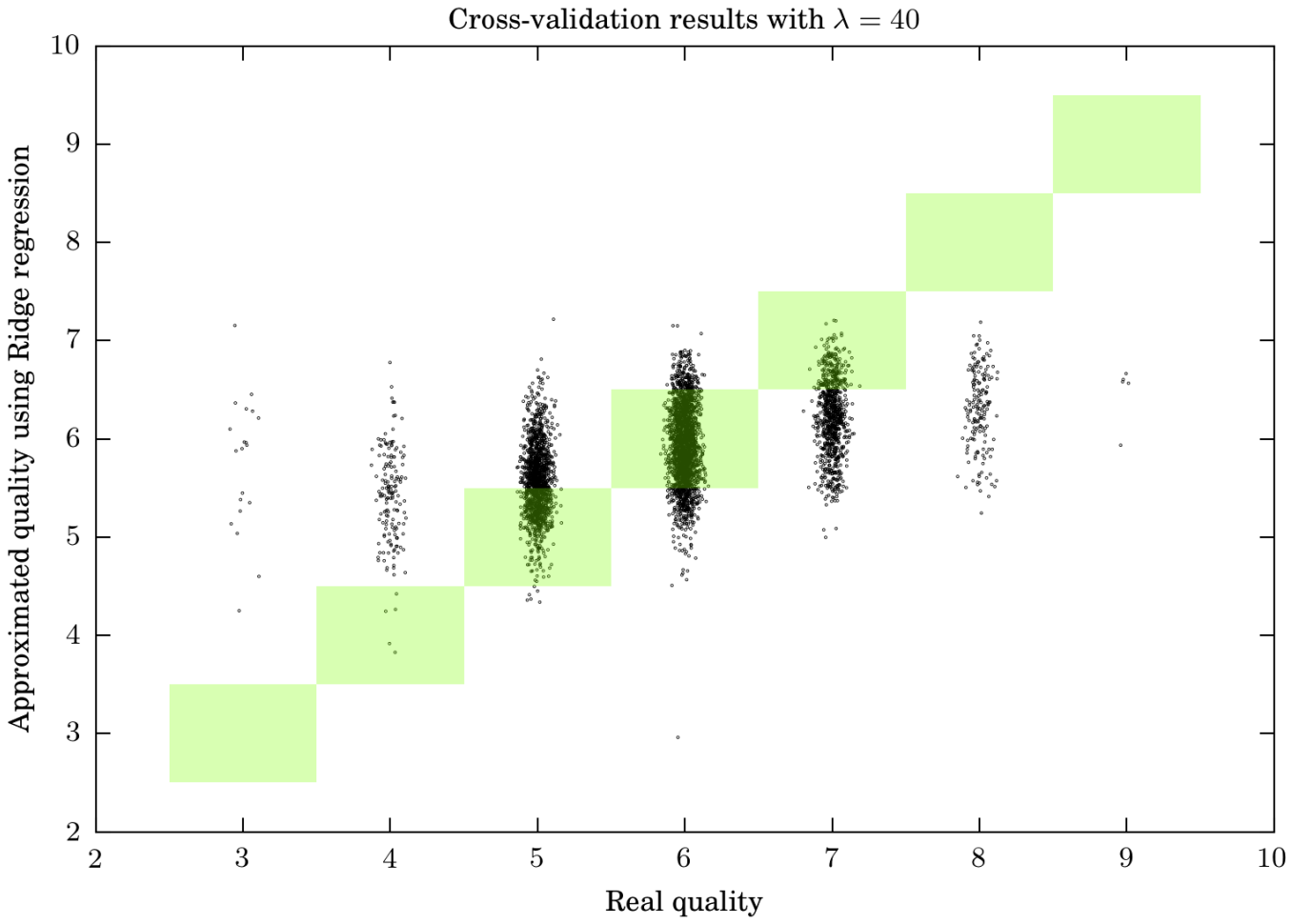

Figure 6: Results of Ridge regression with $\lambda = 40$

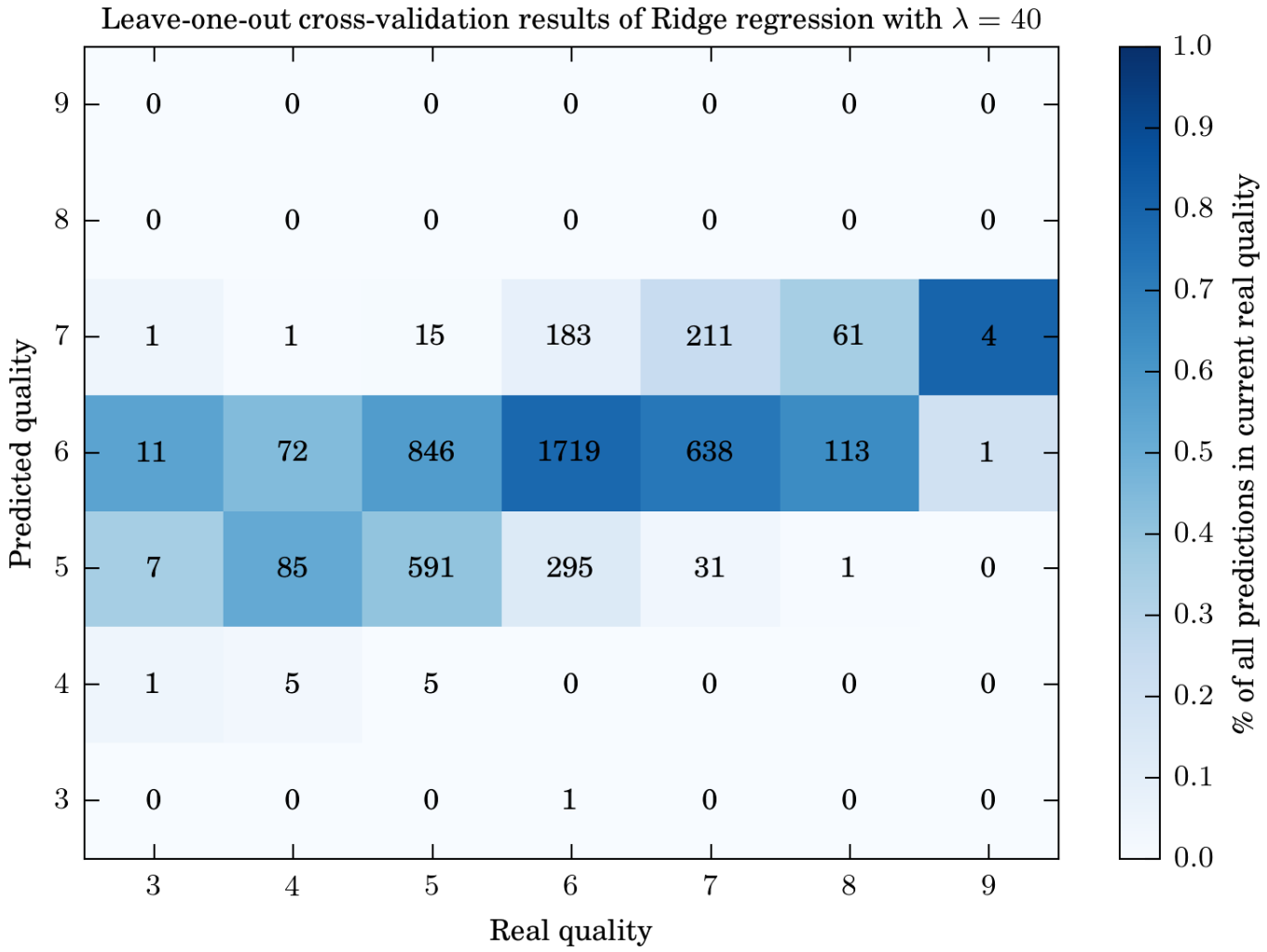

Figure 7: Confusion matrix of Ridge regression cross-validation