Support Vector Regression for Coffee Quality Prediction

Luisa Kalkert

2025-12-27

Support Vector Regression

Data Preprocessing and Cleaning

Feature selection: In SVM Regression, categorical variables are One-Hot Encoded using dummy variables. This means, features with many unique values will lead to a sparse feature matrix. This is disadvantageous as it increases computational cost and poses the risk of overfitting. Additionally, categories with only a few or even only one observation most likely will not add to model robustness, but are adding noise to the signal. Therefore, we are removing categorical features with many unique values, only keeping columns with a low number of unique values. E.g. we are not including the "Region" column with 344 unique values in the model training, because we'd on average have less than 4 observations per Region. For features with a high number of missing values, we either remove the column or impute the missing values. For instance for the "Variety" column, we have about 15% missing values and 8% entries "other" at 30 different varieties. Keeping those, would mean adding 30 sparse columns, and at the same time the information gain is most likely not too high with this many missing values. Therefore, we are not including this feature.

Imputation: For the altitude column, we have about 17% missing values. Here, we are imputing the missing values from the "Region" column. For each observation with missing altitude, we are imputing the median altitude of all coffees from the same region. This covers almost all missing values. For the remaining observations, we are imputing the mean altitude of all coffees from the same country of origin. As the "Region" and "Country" columns are not used in the model training, we are not creating correlated features with this imputation. We are imputing the missing values after the train/test split, using only values from the training set, to avoid data leakage from the test set into the training set.

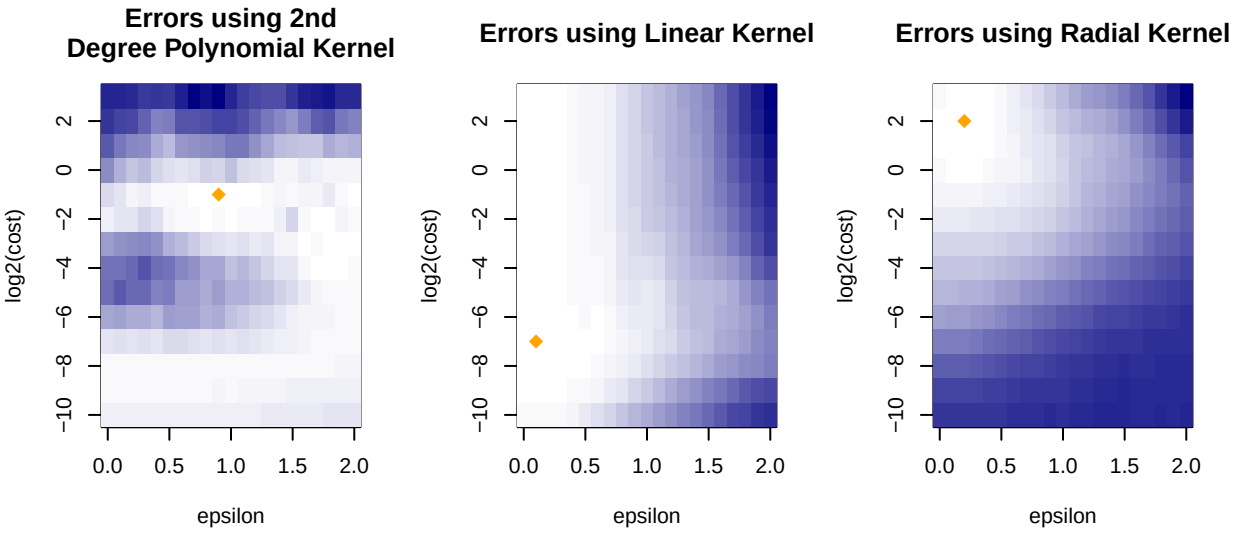
Further preprocessing: The bag weight column contains both pounds and kilograms, so we are converting all to kilograms. However, for some entries it is unclear which unit is used. For these we are averaging the weight in pounds and kilograms. This means for those entries the weight is certainly incorrect, but should be within an order of magnitude of the correct weight. If we can assume that speciality coffee is sold in smaller quantities, we can still gain valuable information from the weight.

Scaling: The SVM method used for model training has a built in scaling functionality. Both predictor and target variables are scaled to have expected value 0 and variance 1. [https://rdrr.io/rforge/e1071/man/svm.html] This transformation is based on the training data and applied to train and test data sets.

Support Vector Regression - Hyperparameter Optimization

For the prediction of cupper points we train a SVM for regression on the data set, holding out the test data set for final evaluation. To choose appropriate hyperparameters, we conduct a hyperparameter optimization on the training data set using 10-fold cross-validation. We use a grid search for different values for cost and

epsilon for a radial, linear and 2-degree polynomial kernel. We compare the mean absolute error between runs, to choose the best performing hyperparameter combination. Finally, we train a model with these hyperparameters on the entire training set and evaluate on the test data set.



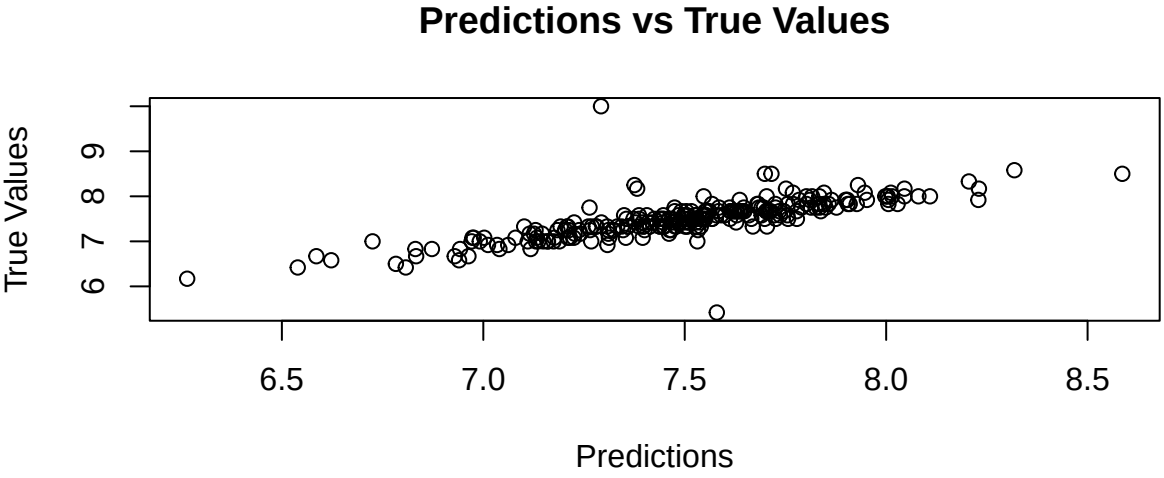
These figures show the mean absolute error for the different hyperparameters by kernel. The best performing hyperparameter combination for each kernel are marked with a orange points.

```
## Best SVM kernel and parameters selected:

##   kernel epsilon      cost     error
## 2 linear     0.1 0.0078125 0.2076482

## [1] "MAE: 0.143099275160558"
```

The best performing kernel is the linear kernel, which means the the relationship between the features and the target variable is best approximated by a multidimensional hyperplane.



Here, we can see the predictions vs. the true values for the test data set. We can see that aside from two outliers, the predictions are close to the true values.