Cross Validation on Polynomial Regression

Hupf Michael

November 20, 2022

Abstract

This document describes the use of cross validation on the *Concrete Compressive Strength* data set to follow as an example and gain insights into cross validation. The results are presented and analyzed.

1 Understanding the data

The used data set *Concrete Compressive Strength* has eight features and a target variable which one wants to predict using polynomial regression. Seven of the eight features are amounts of components which are combined to form concrete. The eighth feature is the age of the concrete in days. A quick glance at the data reveals, that there are no missing values and no values are negative. It's especially important that the values are not negative, because a negative amount is not physically possible.

The target variable is the compressive strength of the concrete, which one can assume depends on the mixture of components.

2 Layout and execution of the experiment

2.1 Data preprocessing

The data is split up into a training (80%) and a test set (20%). The test set will only be used to estimate a final R^2 score once cross validation has found a good model. It's important to note that the test set will not be used during the training process as to avoid overestimating final score.

2.2 Polynomial Regression

The target variable is predicted using linear and polynomial regression, an extension to linear regression where terms of degree 1 to n are also taken into consideration. Polynomial regression tries to fit a regression polynomial instead of a regression line. By transforming the featues $(1, x_1, x_2, ..., x_8)$ one can use a linear regression model to execute polynomial regression. For example on degree two polynomial regression the transformation will produce data of the form $(1, x_1, ..., x_8, x_1^2, ..., x_8^2, x_1x_2, x_1x_3, ..., x_8x_7)$.

The linear models that were used are *Linear Regression* (used linearly here), *Ridge Regression* and *Lasso Regression*. Ridge and Lasso Regression punish absolutely big weights and therefore reduce the chance of overfitting. The amount of punishment is set by a hyperparameter α. The difference between Ridge and Lasso Regression is the used norm, as Ridge uses the euclidian L_2-norm while Lasso Regression uses the L_1-norm. This minor change causes Lasso Regression to reduce weights to exactly zero while Ridge Regression frequently has weights that are approximately zero but not exactly.

2.3 Cross Validation

10-fold cross validation is used to find the hyperparameters n (polynomial degree) and α (regularization parameter). The training data is split up into $k = 10$ more or less equally sized folds. For every iteration from $i = 1, ..., 10$ fold i will be the test set and the other folds the training set. Additionally in every iteration different hyperparameters are used. Afterwards the hyperparameters that yielded the best test score are returned. In this experiment a procedure called random search was used, that means the hyperparameters are randomly selected from a list of options for each iteration.

3 Results

4 Analysis