Linear Regression

Karl Freiwald
Mat-Nr: 3289840

May 31, 2022

Abstract

Linear Regression models linear relationships between continuous input features and a continuous output to predict. The best fit is calculated using all data that is available. Some features seem interfere with the learning process of linear models instead of adding information. We present findings, that demonstrate and substantiate this premonition. The linear models examined in this report lead to the conclusion that features have varying importance in regard to predict. Leaving out features with low significance can enhance the prediction. It does make sense to leave out features and thus use not all information available.

1 Introduction

Linear Regression is probably the most basic algorithm in the field of machine learning. "The core idea is to obtain a line [or hyperplane] that best fits the data. The best fit [...] is the one for which total prediction error (all data points) are as small as possible. " [1] In this report we measure the error using R^2 score, also known as the coefficient of determination.

A common motto in this field of science is: the more data, the more precise a model can be trained. Linear regression can only learn linear relationships and is not capable of learning complex relationships between features. This report deals with the question whether leaving available data out, consequently reducing the complexity of the model, may increase the performance of a linear model. In experiments it is shown that ignoring features does not necessarily deteriorate and seldomly can even enhance the predictive power of the model.

2 Background

One basic assumption of linear regression is that input features and output depend linearly. This barely meets up with reality, but often is a reasonable start make predictions based on data. To retrieve the coefficients of the regression lines or hyperplanes respectively which respresent our prediction functions $\hat{f}$ we used two approaches:

- normal equation $w = (X^T X)^{-1} X^T y$ which delivers a closed form solution for the unique minimum of the residual sum of squares (short: RSS)
- `scikit-learn`'s `LinearRegression`

To rate the regression lines, we used R^2 score which is defined as:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where the total sum of squares $TSS = \sum(y_i - \overline{y})^2$ represents the variability in the data before regression and $RSS = \sum(y_i - \hat{y}_i)^2$ expresses the remaining unexplained variability. Hence R^2 can be looked at as a percentage of variance that can be explained by regression. Generally, $0 \leq R^2 \leq 1$ for X_{train} and for X_{test}. In case that the distribution of X_{test} differs tremendously from X_{train}, R^2 can become negative for X_{test}. A value of X_{test} near 1 would be desirable.

3 Experiments

In the following two experiments we fit linear regression using a selection of features on two datasets:

- single feature using the normal equation
- all features using `scikit-learn`'s `LinearRegression`
- all features leaving one out using `scikit-learn`'s `LinearRegression`

Both datasets contain only continuous features. Every experiment was repeated 50 times to prevent improbable phenomena that could have falsified the results. In each repetition the data was randomly splitted into train (75%) and test set (25%).

3.1 Advertisement Dataset

The Advertisement Dataset consists of 200 datapoints representing a composition of 3 advertisement channels that produce certain sales, which was to predict. The results of the conducted experiments are shown in Table 1.

	Single Feature	All Features	Left Out Feature
TV	59.31	89.07	33.45
Radio	30.41	89.07	61.53
Newspaper	0.86	89.07	89.47

Table 1: R^2 of linear regression in %

Regression lines on single features varied by a wide range, as shown in Figure 1. Predictions using just a single feature performed considerably worse than regression that

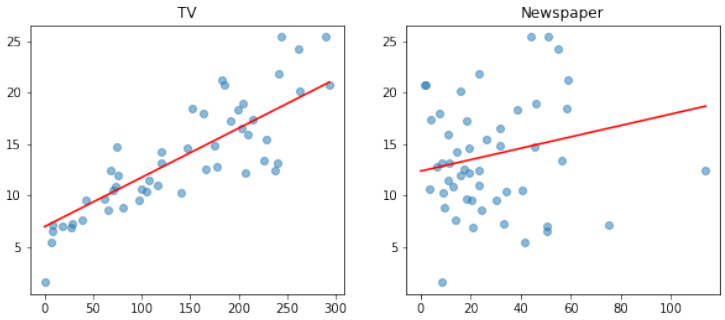

Figure 1: Regression lines on a randomly sampled test data set

takes all features into account. Noticeable is the result of leaving out the feature `Newspaper`. The R^2 score was slightly higher, and thus better, even though less information was used.

3.2 California Housing Dataset

The second experiment was conducted on the California Housing Dataset, which comprises 20640 datapoints with eight features and a `house price` which was the y to predict. Some data have `NaN` values. These were entirely removed, which left a cleansed dataset comprising 20433 examples on which the experiments were conducted.

	Single Feature	All Features	Left Out Feature
Longitude	0.14	63.24	57.48
Latitude	2.00	63.24	56.43
Housing Median Age	1.16	63.24	62.51
Total Rooms	1.77	63.24	63.36
Total Bedrooms	0.19	63.24	63.10
Population	0.02	63.24	61.46
Households	0.37	63.24	63.76
Median Income	47.50	63.24	37.65

Table 2: R^2 of linear regression in %

The only single feature regression that showed a significant R^2 score was `median income`. Left out features barely lowered the predictive power, except for `median income`. Leaving out this feature reduced R^2 from 63.24% down to 37.65%.

4 Conclusion

Single feature linear regression allows barely for any prediction at all. It performs worse than using multiple features in every single experiment. Some features show remarkably high R^2 scores which may indicate a high relevance. Concluding from the results, more features deliver better results, which was to expect. So in general, the common motto "more data, better prediction" holds its truth.

But it is interesting that, apparently, leaving features out in some cases even enhances the precision of prediction. So it can be worthwhile to search out the features that dominate the predictive power and conduct feature selection. A model that uses less features is less complex and therefore can predict more precisely even with a lower number of training samples.[2] A first approach to identify influential features could be when R^2 decreases significantly when this specific feature is left out. Another possible identifier would be when the prediction using solely this feature delivers a reasonable coefficient of determination. Further research could be made on how models behave when only the most significant features are used compared to models that take all available features into account.

References

[1] S. Swaminathan, "Linear regression — detailed view," 2018.

[2] L. Chen, *Curse of Dimensionality.* Springer, Boston, 2009.