**Analysis**

The use of regularization regressions is a deliberate introduction of bias to make the variance of the model lower. This means that the estimated coefficient does not hold much analytical meaning, as they are also biased. For this reason, we will not comment on the particular values of the weight. Rather, the interest and purpose of penalizing model-complexity is to optimize the model’s predictive performance on unseen data. As a measure of performance, we will look at the MSE and furthermore assess the mean absolute errors of the different models as well.

Slide Type-SlideSub-SlideFragmentSkipNotes

Overfitting is: low bias / high variance

* traning our model captures all patterns but we also find some irrelevant
* reacts too much to training sample errors
  + some errors are just noise, and thus we find too many spurious relations
* examples of causes:
  + too much polynomial expansion of variables (PolynomialFeatures)
  + non-linear/logistic without properly tuned hyperparameters:

**Learning Curve:**

We examine the learning curve of the model to examine whether the model is still overfit or if we have found a good bias-variance-balance.

The following plot illustrates the learning curve of our model:

Et billede, der indeholder skærmbillede

Automatisk genereret beskrivelse

The width of the plotted curves expresses the variance of the performance. Interpreting the graph could appear quite daunting, as the performances on the validations set are very inconsistent and varies a lot.

At around 33.000 observations the validation curve drastically decreases in variance, and actually appears quite reasonable. Though the variance is still noteworthy it improves dramatically when the sample gets sufficiently high.

The general behaviour we would expect from a learning curve is this:

* A model of a given complexity will *overfit* a small dataset: this means the training score will be relatively high, while the validation score will be relatively low.
* A model of a given complexity will *underfit* a large dataset: this means that the training score will decrease, but the validation score will increase.
* A model will never, except by chance, give a better score to the validation set than the training set: this means the curves should keep getting closer together but never cross.

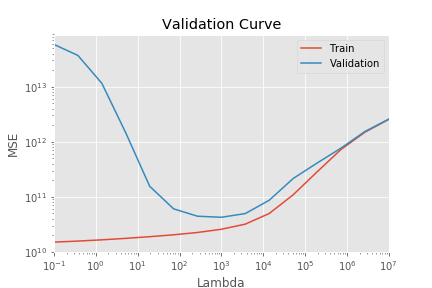
The following figure illustrates the intuition of a learning curve:

Et billede, der indeholder tekst, kort

Automatisk genereret beskrivelse

**Validation Curve:**

We examine the validation curve of the model to examine whether the model is still overfit or if we have found a good bias-variance-balance:



The graph suggests that our model suffers from a degree of overfitting, as it is able to perform remarkably better at predicting on the training data. This is in spite of the fact that we have introduced the Ridge regression to penalize complexity and overfitting.

The following picture illustrates the intuition of a validation curve:

Et billede, der indeholder kort, tekst

Automatisk genereret beskrivelse

Read this, maybe:

<https://jakevdp.github.io/PythonDataScienceHandbook/05.03-hyperparameters-and-model-validation.html>

* For high-bias models, the performance of the model on the validation set is similar to the performance on the training set.
* For high-variance models, the performance of the model on the validation set is far worse than the performance on the training set.

**Conclusion**

In this assignment, we have used machine learning to train a model to predict valuation prices of real estate in Denmark. We did this by scraping Boliga.dk for all of their active houses for sale and combining this with municipal data on income, education, schools and urbanization. Additionally, we also scraped data from hvorlangterder.dk which calculates the distance from an address, to different every-day commodities such as hospitals, schools and shopping.   
We have explained how we carried out our gathering of data, how we have structured the data and how we cleaned it.

We performed descriptive statistics on the data to gain a better acquaintance with the data.

We conducted feature selection by assessing the correlation between the different features and the target variable, which in our assignment has been the valuation price.

We trained both a Lasso- and a Ridge regression model and compared their performances with the performance of a standard linear regression. Though k-fold cross-validation, we estimated the optimized hyperparameters for both regularization models. We found that the Ridge regression performed best on average, with the performance measure being the smallest mean square error of the model.

We have assessed to which degree we have managed to find a good bias-variance-balance for our model and concluded that our model suffers from a degree of overfitting. This was done by examining the validation curve of our final model.

Finally, we have discussed the results of our analysis. We have critically assessed the boundaries of our research and drawn up potential possibilities for improvement.