Abstract

This is the abstract...

Some words on project structure

The main project folder contains the following subfolders:

- data - containing both raw and processed/genereated data used by the programs
- notebooks - containing Jupyter notebooks, representing the "experimental phase" of the project
- reports - containing this final report along with figures and csv files it displays directly
- src - containing Python modules

Although this report summarizes every result obtained, the reader is invited in particular to browse the notebooks, where he or she may follow in detail how the results were obtained. The notebooks represents the gradual evolution of the project in its entirety. Everything is meant to be reproducible - if the reader chooses, he or she can delete the files in `data/generated` and use the notebooks to recreate them. In that case, the following "natural ordering" of the notebooks should be respected:

1. `20190920-Generating-Franke-data.ipynb`
2. `20190905-visualizing-franke.ipynb`
3. `20190906-OLS-Franke.ipynb`
4. `20190918-Resampling-OLS-Franke.ipynb`
5. `20190928-Ridge-regression-Franke.ipynb`

Of course, the reader may also peruse the actual Python modules in the `src` folder. The modules contain unit tests - the testing is briefly described in the special notebook `20190906-testing.ipynb`.

1 Ordinary Least Squares on the Franke Function

In this section, we will generate our target values by sampling the Franke Function on the unit square. This function, which is widely used when testing interpolation and fitting algorithms, is given by

$$
\begin{aligned}
f(x,y) = &\frac{3}{4}\exp\left(-\frac{(9x-2)^2}{4} - \frac{(9y-2)^2}{4}\right) \\
&+ \frac{3}{4}\exp\left(-\frac{(9x+1)^2}{49} - \frac{(9y+1)^2}{10}\right) \\
&+ \frac{1}{2}\exp\left(-\frac{(9x-7)^2}{4} - \frac{(9y-3)^2}{4}\right) \\
&- \frac{1}{5}\exp\left(-(9x-4)^2 - (9y-7)^2\right)
\end{aligned}
$$

It should be noted right away that even if the Franke function is exponential in x and y, Taylor's theorem should guarantee that we will be able to get quite close to the original Franke data by using fifth-degree polynomials. In addition to the "pure" Franke values given by the function, we will generate a couple of perturbed variations by adding normally distributed noise. This will give us the opportunity to study how it will be more difficult to create a good fit to the data as it grows more complex.

1.1 Generating and visualizing data

The Franke data are generated in the notebook `20190920-Generating-Franke-data.ipynb`, which relies on `src/data/generate_data.py`. By varying the noise term, we get three sets of target values, which are stored in separate files:

- `no_noise.csv` with no noise added to the Franke data
- `some_noise.csv`, where normally distributed noise with mean 0 and standard deviation 0.1 is added to the Franke data
- `noisy.csv`, where normally distributed noise with mean 0 and standard deviation 0.4 is added to the Franke data

The data are stored in `data/generated/`. Of course, the data could easily have been generated on the fly as needed, but by storing it at this point, we facilitate reproducibility, as we do not run the risk of generating lots of datasets that may look similar, but actually are different from each other. From this point on, we will work with the data read from the files, and nothing else.

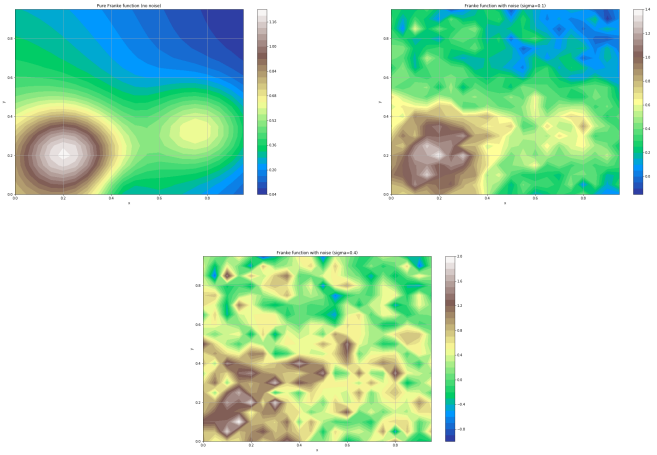
In the same notebook (`20190920-Generating-Franke-data.ipynb`) we also generate a feature matrix X. In addition to the original grid points $\{(x = 0.05i, y = 0.05j) : i, j = 0, ..., 19\}$ we include features of the form $x^m y^n$, where m and n are non-negative integers and $m + n \leq 5$. This is in order to perform polynomial regression analysis, which is just another name for ordinary linear regression performed on a feature matrix augmented by polynomial combinations of the original features. The heavy lifting in constructing the polynomial features is performed by the class `PolynomialFeatures` in `src/features/polynomial.py`. Again, we choose to store the generated feature matrix for subsequent use (`data/generated/X.csv`).

We now have one common feature matrix and three different sets of target values. In the notebook `20190905-visualizing-franke.ipynb`, we perform some simple exploratory data analysis on those datasets. The table below shows the output of the `describe` function from `pandas`[1].

	No noise	Some noise (sigma 0.1)	Noisy (sigma 0.4)
count	400	400	400
mean	0.43	0.43	0.42
std	0.28	0.3	0.49
min	$4.5 \cdot 10^{-2}$	-0.13	-0.93
25	0.23	0.21	$8.64 \cdot 10^{-2}$
50	0.35	0.37	0.42
75	0.57	0.59	0.71
max	1.22	1.37	1.96

The table gives us some insight into the difference between the three sets of target values, in particular when we look at the spread (compare the standard deviation and the difference between the max and min rows.)

We can also get a feel of the data by plotting them. The code for the following contour plots is in the notebook `20190905-visualizing-franke.ipynb`. We see that the original Franke function seems quite smooth, whereas the noisy version is full of "spikes" that will make fitting harder.



[1] I have saved the output into the file `1_description_table.csv` and loaded it into LaTeX using the `pgfplotstabletypeset` package, but as my I am not able to display the percentage signs in the row names of the percentile rows.

1.2 Ordinary Least Squares regression - theoretical recap

1.2.1 Function fitting

We now proceed to actually fitting functions to the data points. The OLS (Ordinary Least Squares) method is very well known, so the following description will be brief. Given data points $\{(x_i, y_i) : i = 1, ..., m\}$, where each x_i is a vector in $\mathbb{R}^n$ and each y_i is a real number, we want to obtain a parameter vector $\hat{\beta} \in \mathbb{R}$ so that for each input vector x_i, the predicted output $\hat{y_i} = {x_i}^T \hat{\beta}$ will be "close to" the target value y_i. One possible way to do this is to choose $\hat{\beta}$ so as to minimize the sum of squared errors, that is, to set

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{m} (y_i - \hat{y_i})^2$$

If we collect the (x_i) in a feature matrix

$$\mathbf{X} = \begin{bmatrix} {x_1}^T \\ {x_2}^T \\ .... \\ {x_m}^T \end{bmatrix}$$

we easily see that the above minimization problem can be formulated as

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2$$

The minimization problem turns out to have a closed form solution, which can be found using matrix differentiation. The solution is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Of course, this solution only exists if $\mathbf{X}^T\mathbf{X}$ is nonsingular, but this is rarely a problem in real-life situations. There are more effective ways of computing the least-squares estimate for β, but we stick to matrix inversion here, as it is very easy to implement. The implementation is in the `fit` method of the `OLS` class in `src/models/models.py`.

1.2.2 Estimates of error and variability

Once we have obtained our estimates for β, we can get predictions $\hat{\mathbf{y}}$ by putting $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$. We then can assess the quality of our estimates in various ways. The MSE (Mean Square Error) is one such measure. Given target values $\mathbf{y}$ and predictions $\hat{\mathbf{y}}$, the MSE is given by $\frac{1}{m}\|\mathbf{y} - \hat{\mathbf{y}}\|^2$. The implementation of the MSE is in `src/evaluation/evaluation.py`, along with a function for evalutating the R^2 score. The R^2 score measures how much the error is reduced by using our linear model as opposed to a pure mean model, that is, a model where we predict $\hat{y_i} = \frac{1}{m}\sum_{i=1}^{m} y_i$ for $i = 1, ..., m$. The ratio between the mean square errors of the two models is

$$\frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{\mathbf{y}})^2}$$

where $\bar{\mathbf{y}} = \frac{1}{m}\sum_{i=1}^{m} y_i$. We want this ratio to be low, that is, we want our model predictions to be much better than the values predicted by the pure mean model. This leads us to defining an measure

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{\mathbf{y}})^2}$$

which will hopefully be close to 1.

The last measure we are going to discuss is the variance of our estimates for β. This is an important issue, because when doing linear regression, we assume a "ground truth" model $y_i = {x_i}^T\beta + \epsilon$, where the ϵ is a normally distributed random variable. This means that $\hat{\beta}$ is a random variable too, and we are interested in its variance, which will allow us to construct condifence intervals around our point estimates. It can be shown that the covariance matrix of $\hat{\beta}$ is $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, where σ^2 is the variance of ϵ. Because we are mainly interested in the variance of the individual components of $\hat{\beta}$, we focus on the diagonal of this matrix, and find that $\mathrm{Var}(\hat{\beta}_j) = \sigma^2((\mathbf{X}^T\mathbf{X})^{-1})_{j,j}$. In general, σ^2 will be unknown to us, but it can be estimated by the MSE we calculated previously. This leads to the expression

$$\hat{\sigma}^2(\hat{\beta}_j) = \mathrm{MSE} \times ((\mathbf{X}^T\mathbf{X})^{-1})_{j,j}$$

as our estimate of the variance of ${\hat{\beta}_j}^2$ Having obtained variance estimates, we can take their square root $\hat{\sigma}_{\hat{\beta}_j}$ and construct 95 % confidence intervals as

$$\left[\hat{\beta}_j - 1.96\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + 1.96\hat{\sigma}_{\hat{\beta}_j}\right]$$

1.3 Ordinary Least Squares regression - results

In the notebook `20190906-OLS-Franke.ipynb`, we perform OLS regression on our generated datasets and evaluate our results using the metrics described above. This notebook depends upon `src/models/models.py` for the actual data fitting and `src/evaluation/evaluation.py` for the metrics. We begin by comparing the output from the regression method with the target values of the three datasets, and obtain the following results:

	MSE	R^2
No noise	$2.1434 \cdot 10^{-3}$	0.97202
Some noise (sigma 0.1)	0.01064	0.87914
Noisy (sigma 0.4)	0.15188	0.36873

[2]Have I misunderstood anything here? According to the lecture note, the square root of $((\mathbf{X}^T\mathbf{X})^{-1})_{j,j}$ should be taken in this formula, but this does not seem right.

As we see, the R^2 scores quite close to one in the pure Franke case, meaning that our model explains most of the variability in the data. In the more noisy datasets, the R^2 is much worse. The MSE, as expected, grows significally with added noise. A fifth-degree polynomial can be fitted rather well to a pure Franke function, whereas with added noise, the fitting becomes much harder.

We now proceed to the actual parameter estimates and their variances. Because we have three datasets and there are 21 parameters, we get quite a lot of data here. In the notebook `20190906-OLS-Franke.ipynb` we construct three csv files corresponding to the three datasets. For the no-noise case, we get the following data:

Feature	Estimate	Variance	Lower bound	Upper bound
1	0.54154	$7.55532 \cdot 10^{-4}$	0.48766	0.59541
x	6.07027	0.12238	5.3846	6.75593
y	3.16585	0.12238	2.48018	3.85151
(x^2)	−27.6196	3.4886	−31.28044	−23.95875
(y)(x)	−11.6201	2.01247	−14.40059	−8.83961
(y^2)	−6.30394	3.4886	−9.96479	−2.6431
(x^3)	32.72699	20.48258	23.85648	41.5975
$(y)(x^2)$	41.72654	10.53252	35.36559	48.08749
$(y^2)(x)$	18.58331	10.53252	12.22236	24.94427
(y^3)	−15.79604	20.48258	−24.66655	−6.92553
(x^4)	−3.94401	25.24997	−13.79288	5.90486
$(y)(x^3)$	−60.43109	13.81827	−67.71699	−53.1452
$(y^2)(x^2)$	0.22793	11.68392	−6.47169	6.92756
$(y^3)(x)$	−33.43817	13.81827	−40.72407	−26.15228
(y^4)	41.04946	25.24997	31.20058	50.89833
(x^5)	−7.93131	4.30853	−11.99968	−3.86293
$(y)(x^4)$	25.99715	3.06771	22.56423	29.43006
$(y^2)(x^3)$	5.77911	2.83309	2.48008	9.07814
$(y^3)(x^2)$	−5.34619	2.83309	−8.64522	−2.04716
$(y^4)(x)$	19.13743	3.06771	15.70451	22.57034
(y^5)	−22.59142	4.30853	−26.6598	−18.52305

This table shows the β estimates for each individual feature, along with the estimates for the variance and the lower and upper bound for the corresponding confidence intervals.

For the case where normally distributed noise with $\sigma = 0.1$ is added to the output from the Franke function, we get the following parameter estimates and confidence intervals:

Finally, for the case where normally distributed noise with $\sigma = 0.4$ is added, we get this table:

Feature	Estimate	Variance	Lower bound	Upper bound
1	0.57132	$3.74901 \cdot 10^{-3}$	0.45131	0.69133
x	6.28238	0.60726	4.75501	7.80975
y	2.8003	0.60726	1.27293	4.32768
(x^2)	−28.66944	17.31069	−36.82423	−20.51464
(y)(x)	−13.88187	9.98605	−20.07561	−7.68814
(y^2)	−3.36449	17.31069	−11.51929	4.79031
(x^3)	33.98078	101.6362	14.22109	53.74048
$(y)(x^2)$	49.92817	52.26317	35.75869	64.09765
$(y^2)(x)$	21.24688	52.26317	7.0774	35.41636
(y^3)	−24.48217	101.6362	−44.24187	−4.72247
(x^4)	−4.49262	125.29237	−26.4317	17.44645
$(y)(x^3)$	−69.94827	68.56736	−86.17813	−53.71841
$(y^2)(x^2)$	−4.23323	57.97655	−19.15712	10.69067
$(y^3)(x)$	−35.41406	68.56736	−51.64392	−19.1842
(y^4)	51.63963	125.29237	29.70055	73.57871
(x^5)	−7.76393	21.37928	−16.82652	1.29867
$(y)(x^4)$	29.27725	15.22221	21.63018	36.92432
$(y^2)(x^3)$	8.55686	14.05803	1.20803	15.90569
$(y^3)(x^2)$	−4.82208	14.05803	−12.17091	2.52675
$(y^4)(x)$	19.95001	15.22221	12.30295	27.59708
(y^5)	−27.15941	21.37928	−36.22201	−18.09682

Feature	Estimate	Variance	Lower bound	Upper bound
1	0.66068	0.05353	0.20718	1.11418
x	6.91873	8.67158	1.14701	12.69045
y	1.70367	8.67158	−4.06805	7.47539
(x^2)	−31.81895	247.19257	−62.63477	−1.00312
(y)(x)	−20.66721	142.59836	−44.07246	2.73805
(y^2)	5.45385	247.19257	−25.36197	36.26968
(x^3)	37.74215	1,451.34072	−36.92692	112.41123
$(y)(x^2)$	74.53306	746.30569	20.98861	128.07751
$(y^2)(x)$	29.23759	746.30569	−24.30686	82.78204
(y^3)	−50.54057	1,451.34073	−125.20964	24.12851
(x^4)	−6.13847	1,789.14529	−89.04312	76.76618
$(y)(x^3)$	−98.49979	979.12557	−159.83011	−37.16946
$(y^2)(x^2)$	−17.61671	827.89135	−74.01198	38.77857
$(y^3)(x)$	−41.34171	979.12557	−102.67203	19.98862
(y^4)	83.41016	1,789.14529	0.50552	166.31481
(x^5)	−7.26178	305.29107	−41.50804	26.98447
$(y)(x^4)$	39.11755	217.36947	10.22038	68.01473
$(y^2)(x^3)$	16.89011	200.74525	−10.88007	44.66029
$(y^3)(x^2)$	−3.24976	200.74525	−31.01994	24.52042
$(y^4)(x)$	22.38777	217.36947	−6.50941	51.28494
(y^5)	−40.86339	305.29107	−75.10965	−6.61713

2 Resampling techniques, adding more complexity

Hitherto, we have evaluated our results using the same data we used for model fitting. In most real-life scenarios, this is considered bad practice, as the models tend to "learn" noise in addition to the underlying distribution of the data. Thus, we easily end up with models that are able to reproduce the original data quite well, but do not necessarily deal equally well with "unknown" data. This is known as the overfitting problem.

In order to get a better picture of how well our models generalize to unseen data, we use different resampling techniques. We can for instance split our data into two parts, known as the test set and the training set. We use the training set for model fitting and the test set for evaluation. As the test set consists entirely of unseen data, this gives us a much better understanding of the generalizing ability of the model.

In the notebook `20190918-Resampling-OLS-Franke.ipynb` we use the function `train-test-split` from `sklearn.model_selection` in order to achieve such a split. When evaluating our model's ability to reproduce the test data after having been trained on the training data, we get the following MSE and R^2:

	MSE	R^2
No noise	$3.42806 \cdot 10^{-3}$	0.95562
Some noise (sigma 0.1)	0.01482	0.82235
Noisy (sigma 0.4)	0.20929	0.11503

If we compare this table to the corresponding table we made earlier, we see that the MSE is higher, in particular in the noisy case (0.209 versus 0.152). This is of course to be expected, as we now test our model on unseen data. The R^2 scores visibly lower than before.

Resampling can be done in other ways as well. One possible approach is k-fold cross-validation, where we start by partitioning our data $\mathcal{D}$ into k disjoint folds $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_k$. For $j = 1, ..., k$, we then use $\mathcal{D} \setminus \mathcal{D}_j$ as our training set and $\mathcal{D}_j$ as our test set. For each value of j, we can obtain a MSE score, and by averaging the MSE scores, we hopefully get a precise estimate of the true MSE.

In the same notebook (`20190918-Resampling-OLS-Franke.ipynb`) we use the class `K_fold_splitter` from `src/resampling/resampling.py` to estimate the test MSE using cross-validation. We obtain the following results on our three datasets:

	MSE
No noise	$2.64309 \cdot 10^{-3}$
Some noise (sigma 0.1)	0.01235
Noisy (sigma 0.4)	0.17156

3 Bias-variance tradeoff

Above we mentioned the importance of validating our models on unseen data in order to avoid overfitting. In this section, we will bring this somewhat "intuitive" idea onto a more precise theoretical footing. In particular, we will decompose our test MSE into three separate terms, of which the first may be interpreted as error stemming from adopting too rigid a model, which may not be able to capture the relevant fluctuations in the training set properly. This is called the bias (or to be more precise, the square of the bias). An obvious way to reduce the bias is to adopt a more flexible model, but this may easily introduce a second kind of error, which is the one described above, i. e. that the model fits too closely to the training set, picking up fluctuations due to sampling rather than to the true underlying distribution of the data. This second kind of error is known as variance.

As stated above, we take as our point of departure a dataset $\{(x_i, y_i) : i = 1, ..., m\}$ and a "ground truth" model $\mathbf{y} = f(x) + \epsilon$. Hitherto, the function f has been of the linear form $f(x) = x^T\beta$, but in the following discussion we will consider general models of this type. What is important is that the noise term ϵ is normally distributed with mean 0 and variance σ^2.

When doing regression analysis, our goal is to obtain a model $\hat{f}$ to obtain predictions $\hat{\mathbf{y}} = \hat{f}(x)$. In this setting, both $\mathbf{y}$, $\hat{\mathbf{y}}$ and ϵ are random variables. We have already stated that $\mathbb{E}(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2$, and now proceed to compute the expectations and variances of $\mathbf{y}$ and $\hat{\mathbf{y}}$. To unclutter notation, we put $f = f(x)$, and note that since f is deterministic, its expected value is f and its variance is zero. We have

$$\begin{aligned}\mathbb{E}(\mathbf{y}) &= \mathbb{E}(f + \epsilon) \\ &= \mathbb{E}(f) + \mathbb{E}(\epsilon) \\ &= f\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}(\mathbf{y}) &= \mathbb{E}[(\mathbf{y} - \mathbb{E}(\mathbf{y}))^2] \\ &= \mathbb{E}[(f + \epsilon - f)^2] \\ &= \mathbb{E}[(\epsilon - 0)^2] \\ &= \mathrm{Var}(\epsilon) \\ &= \sigma^2\end{aligned}$$

When doing regression analysis, we try to minimize the test MSE, which can be written as $\mathbb{E}[(\mathbf{y} - \hat{\mathbf{y}})^2]$. We can rewrite this in the following way. We set $u = \mathbb{E}(\hat{\mathbf{y}})$, again in an attempt to unclutter notation somewhat.

$$
\begin{aligned}
\mathbb{E}[(\mathbf{y}-\hat{\mathbf{y}})^2] &= \mathbb{E}[(f+\epsilon-\hat{\mathbf{y}})^2] \\
&= \mathbb{E}[(f+\epsilon-\hat{\mathbf{y}}+u-u)^2] \\
&= \mathbb{E}[((f-u)+\epsilon+(u-\hat{\mathbf{y}}))^2] \\
&= \mathbb{E}[(f-u)^2+\epsilon^2+(u-\hat{\mathbf{y}})^2+2\epsilon(f-u)+2\epsilon(u-\hat{\mathbf{y}})+2(f-u)(u-\hat{\mathbf{y}})] \\
&= \mathbb{E}[(f-u)^2]+\mathbb{E}[\epsilon^2]+\mathbb{E}[(u-\hat{\mathbf{y}})^2]+\mathbb{E}[2\epsilon(f-u)]+\mathbb{E}[2\epsilon(u-\hat{\mathbf{y}})]+\mathbb{E}[2(f-u)(u-\hat{\mathbf{y}})] \\
&= (f-u)^2+\sigma^2+\mathrm{Var}(\hat{\mathbf{y}})+0+0+0 \\
&= (f-\mathbb{E}(\hat{\mathbf{y}}))^2+\mathrm{Var}(\hat{\mathbf{y}})+\sigma^2 \\
&= \frac{1}{m}\sum_1^m (f_i-\mathbb{E}(\hat{\mathbf{y}}))^2+\frac{1}{m}\sum_1^m (\hat{y}_i-\mathbb{E}(\hat{\mathbf{y}}))^2+\sigma^2
\end{aligned}
$$

The first of these terms is the bias, which can be explained as the error resulting from erroneous assumptions in the learning algorithm, for instance that we assume a linear model when the actual underlying distribution is non-linear. The second term is the variance, which is the error resulting from our model picking up and incorporating randomness in the training set due to sampling. The last term is the irreducible error, stemming from the fact that the true underlying model $\mathbf{y} = f(x) + \epsilon$ has a random noise term.

As the complexity of our model increases, it will more easily adapt to fluctuations in the training set. Thus while the bias decreases, the variance usually increases. This leads to an interesting trade-off situation: If we choose too simple a model, we will not be able to pick up important underlying patterns in the data, and thus increasing our test MSE. However, if we make our model too complex, it will model noise in the training data, again increasing the test MSE.

In the notebook `20191002-Training-versus-test-error-plot-OLS.ipynb`, we try out polynomial models of degrees 0, 1, 2, 3, 4 and 5 on the noisy dataset. Thus we increase model complexity, and should expect a monotonically decreasing train MSE, whereas the test MSE could in principle increase. We obtain the following plot:

We clearly see that in this particular case, we seem to reach a "sweet spot" around degree 3, where the test MSE is optimal. Of course, further testing should be conducted in order to find out if this is a general tendency.

4 Ridge regression on the Franke function with resampling

As described above, we have fitted our OLS model parameters by minimizing the sum of squared errors on the training set. That is, we have solved the minimization problem

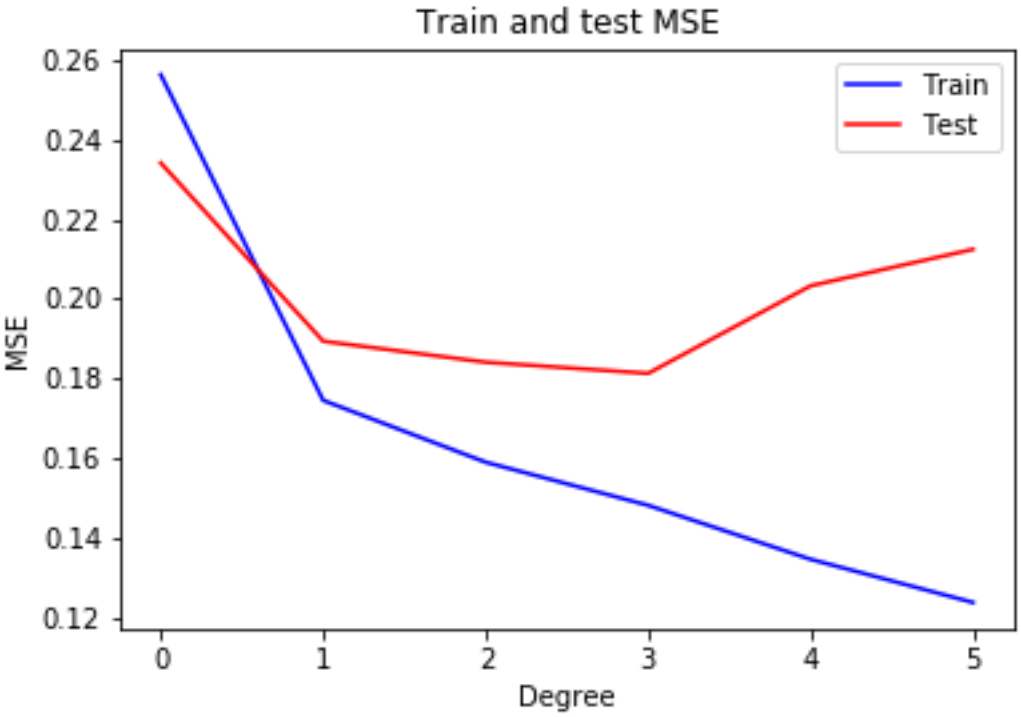

Train and test MSE
Train
Test
MSE
0.26
0.24
0.22
0.20
0.18
0.16
0.14
0.12
0
1
2
3
4
5
Degree

Figure 2

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2$$

The idea behind ridge regression is to introduce a penalty term that will "dampen" the obtained parameters somewhat by bringing them closer to zero. The minimization problem now becomes

$$\hat{\beta} = \arg\min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda \|\beta\|^2$$

Here, λ is a non-negative real number. If $\lambda = 0$, we get our usual OLS cost function. Higher values of λ corresponds to more significant "damping" or regularization. As we have already stated that our original OLS approach has a closed-form solution that minimizes the sum of squared errors on the training set, we cannot expect our new approach to obtain a better result on the training set. This, however, is not really a problem, since we are interested in are models that perform well on unseen data, not on the same data as it was trained on. Our hope is that as we increase the bias slightly by forcing the β parameters closer to zero, we reduce the variance so that the model will be less prone to overfitting. Hopefully, this will lead to a lower test MSE.

It can be shown, using matrix differentiation, that the Ridge regression minimization problem has a closed-form solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

The implementation in `src/models/models.py` is based on this equation. It should be noted that the situation is much more complex here than with Ordinary Least Squares, since we obtain a multitude (actually, an infinity) of models depending on the hyperparameter λ.

In the notebook `20190928-Ridge-regression-Franke.ipynb`, we try out Ridge regression on our three datasets. (We return to working only with fifth-degree polynomials here.) We use cross-validation to search for the value of λ that yields the lowest test MSE. We discover that it is only the noisy dataset (the one where we added normally distributed noise with standard deviation 0.4 to the Franke function) that actually benefits from Ridge regression. For the two other datsets, the optimal λ hyperparameter found is 0, meaning that ordinay least squares actually works best.

For the noisy dataset, however, we find that a λ parameter of 0.0101 yields a test MSE of 0.17080, which is slightly lower than 0.17156, which is the test MSE obtained by letting λ equal 0. The difference may be insignificant, but a tentative conclusion could be that the presence of noise increases the risk of overitting somewhat, so that a little regularization may be a good idea. A plot of test MSE versus lambda on the noisy dataset seems to confirm this:

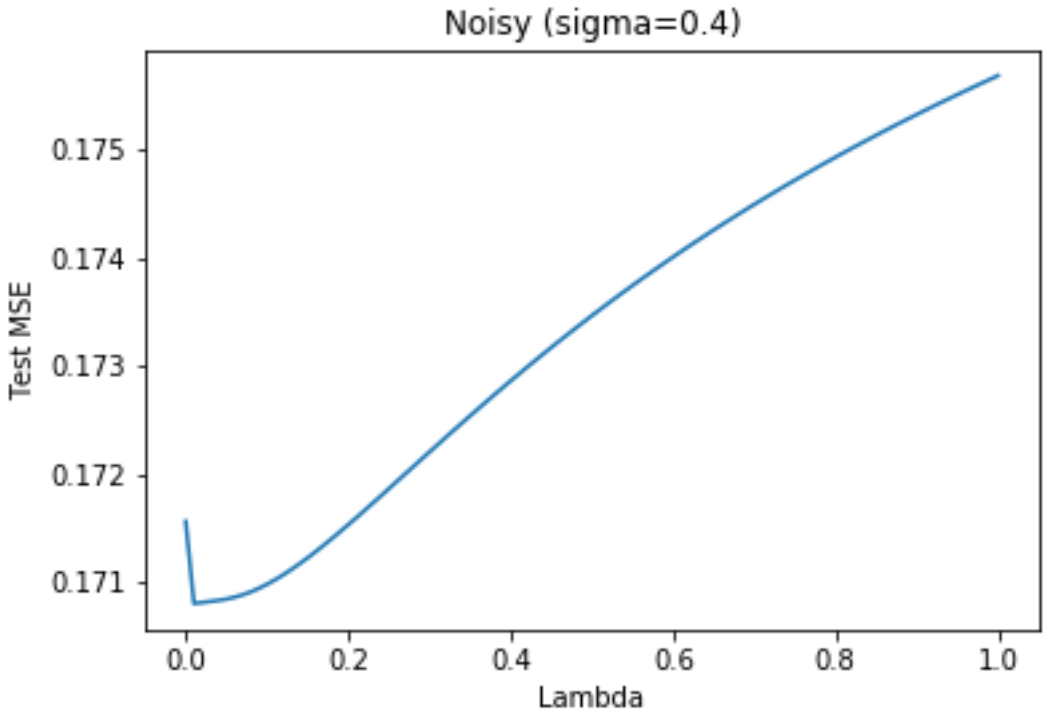


Figure 3

5 Lasso regression on the Franke function with resampling

6 Real data

After having tested our methods for fitting and evaluation on generated datasets, we are now hopefully ready to work with real data.