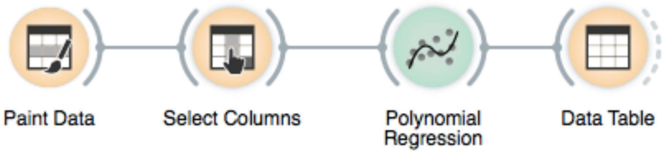
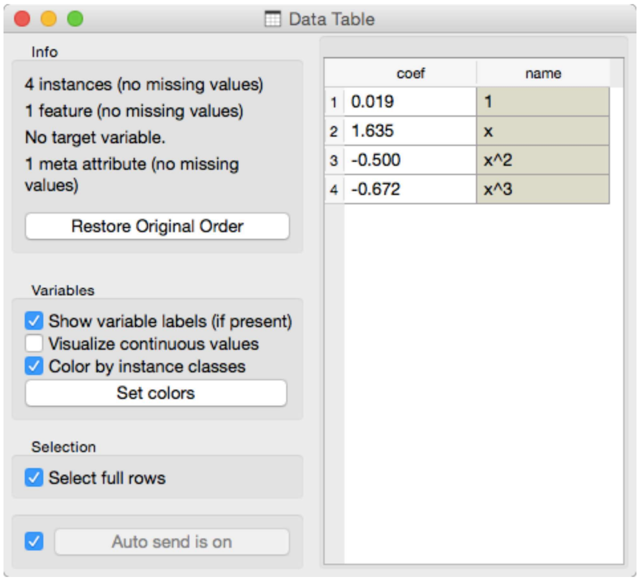
Lesson 18: Regularization

There has to be some cure for the overfitting. Something that helps us control it. To find it, let's check what the values of the parameters θ under different degrees of polynomials actually are



With smaller degree polynomials values of θ stay small, but then as the degree goes up, the numbers get really large.



Data Table

Info

4 instances (no missing values)
1 feature (no missing values)
No target variable.
1 meta attribute (no missing values)

Restore Original Order

Variables

☑ Show variable labels (if present)
☐ Visualize continuous values
☑ Color by instance classes

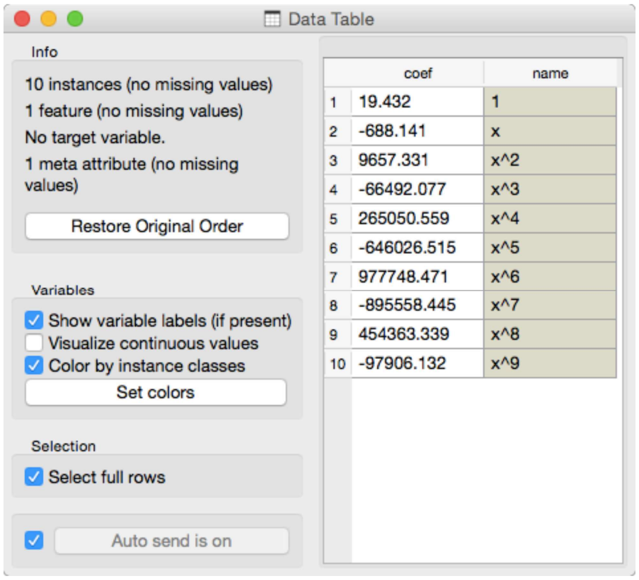
Set colors

Selection

☑ Select full rows

☑ Auto send is on

	coef	name
1	0.019	1
2	1.635	x
3	-0.500	x^2
4	-0.672	x^3



Data Table

Info

10 instances (no missing values)
1 feature (no missing values)
No target variable.
1 meta attribute (no missing values)

Restore Original Order

Variables

☑ Show variable labels (if present)
☐ Visualize continuous values
☑ Color by instance classes

Set colors

Selection

☑ Select full rows

☑ Auto send is on

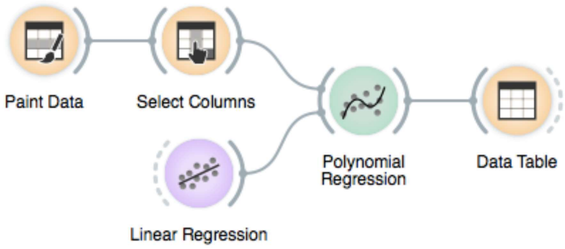
	coef	name
1	19.432	1
2	-688.141	x
3	9657.331	x^2
4	-66492.077	x^3
5	265050.559	x^4
6	-646026.515	x^5
7	977748.471	x^6
8	-895558.445	x^7
9	454363.339	x^8
10	-97906.132	x^9

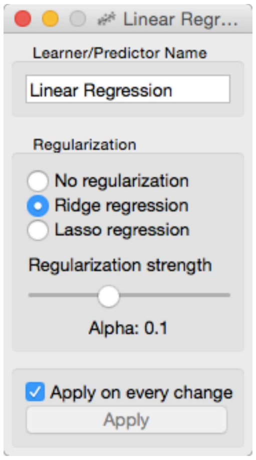
More complex models can fit the training data better. The fitted curve can wiggle sharply. The derivatives of such functions are high, and so need to be the coefficients θ. If only we could force the linear regression to infer models with a small value of coefficients. Oh, but we can. Remember, we have started with the optimization function the linear regression minimizes, the sum of squared errors. We could simply add to this a sum of all θ squared. And ask the linear regression to minimize both terms. Perhaps we should weigh the part with θ squared, say, we some coefficient λ, just to control the level of regularization.

Which inference of linear model would overfit more, the one with high λ or the one with low λ? What should the value of λ be to cancel regularization? What if the value of λ is really high, say 1000?

Here we go: we just reinvented regularization, a procedure that helps machine learning models not to overfit the training data. To observe the effects of the regularization, we can give Polynomial Regression our own learner, which supports these kind of settings.

Internally, if no learner is present on its input, the Polynomial Regression widget would use just its ordinary, non-regularized linear regression.




Linear Regr...

Learner/Predictor Name

Linear Regression

Regularization

○ No regularization
● Ridge regression
○ Lasso regression

Regularization strength

Alpha: 0.1

☑ Apply on every change

Apply

The Linear Regression widget provides two types of regularization. Ridge regression is the one we have talked about and minimizes the sum of squared coefficients θ. Lasso regression minimizes the sum of absolute value of coefficients. Although the difference may seem negligible, the consequences are that lasso regression may result in a large proportion of coefficients θ being zero, in this way performing feature subset selection.

Now for the test. Increase the degree of polynomial to the max. Use Ridge Regression. Does the inferred model overfit the data? How does degree of overfitting depend on regularization strength?