Training, validating and testing

Herman Kamper

http://www.kamperh.com/

How do we know which model to pick?

E.g. for linear regression we can choose the degree of polynomial features, the number of RBF basis functions, and the value of λ for L_1 or L_2 regularisation. How do we choose which of these to use?

Classification: e.g. accuracy

We could look at the loss (or another metric such as MSE or RMSE) on the training data. But this is problematic since the more complex/expressive model (e.g. high order polynomial, lower values of λ) will always do better (on the training data).

Idea: Use a held-out validation set (which you don't train on) and evaluate different model options on that.

$$\underbrace{(\underline{x}^{(1)}, y^{(1)}), (\underline{x}^{(2)}, y^{(2)}), \ldots}_{\text{Training set}} \underbrace{(\underline{x}^{(N)}, y^{(N)})}_{\text{Validation set}}$$

80% — Fit models on this part

20% — Evaluate on this part

How well does the model generalize?

How well could you expect the model to perform on completely new data?

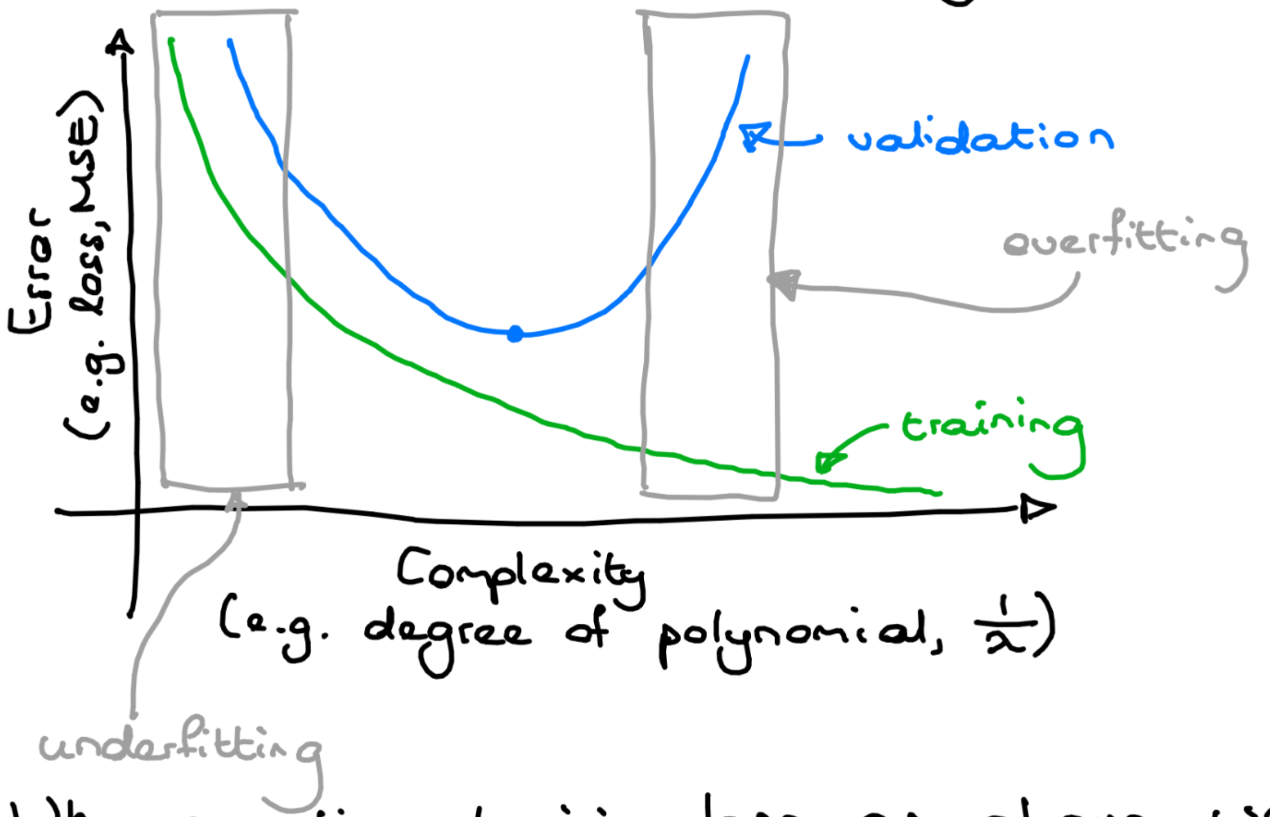
Could report validation metric, but this is likely to be optimistic (we actually had to fit some hyperparameters like λ on the validation data).

Idea: Report final performance on a completely held-out test set. Try to use this set as little as possible (ideally never) in model development.

$$\boxed{(\underline{x}^{(1)}, y^{(1)}),\ (\underline{x}^{(2)}, y^{(2)}),\ \ldots\ (\underline{x}^{(N)}, y^{(N)})}$$

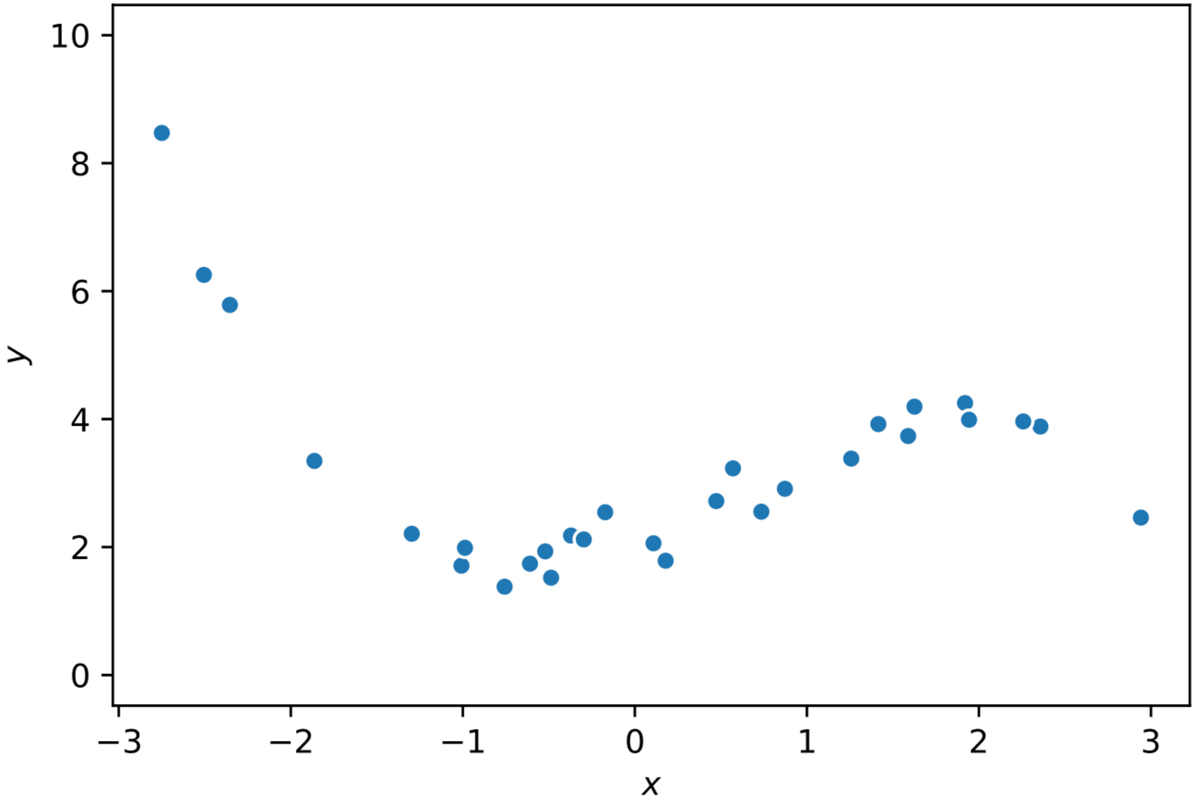
Training	Validation	Testing
80%	10%	10%

What does over- and underfitting look like?

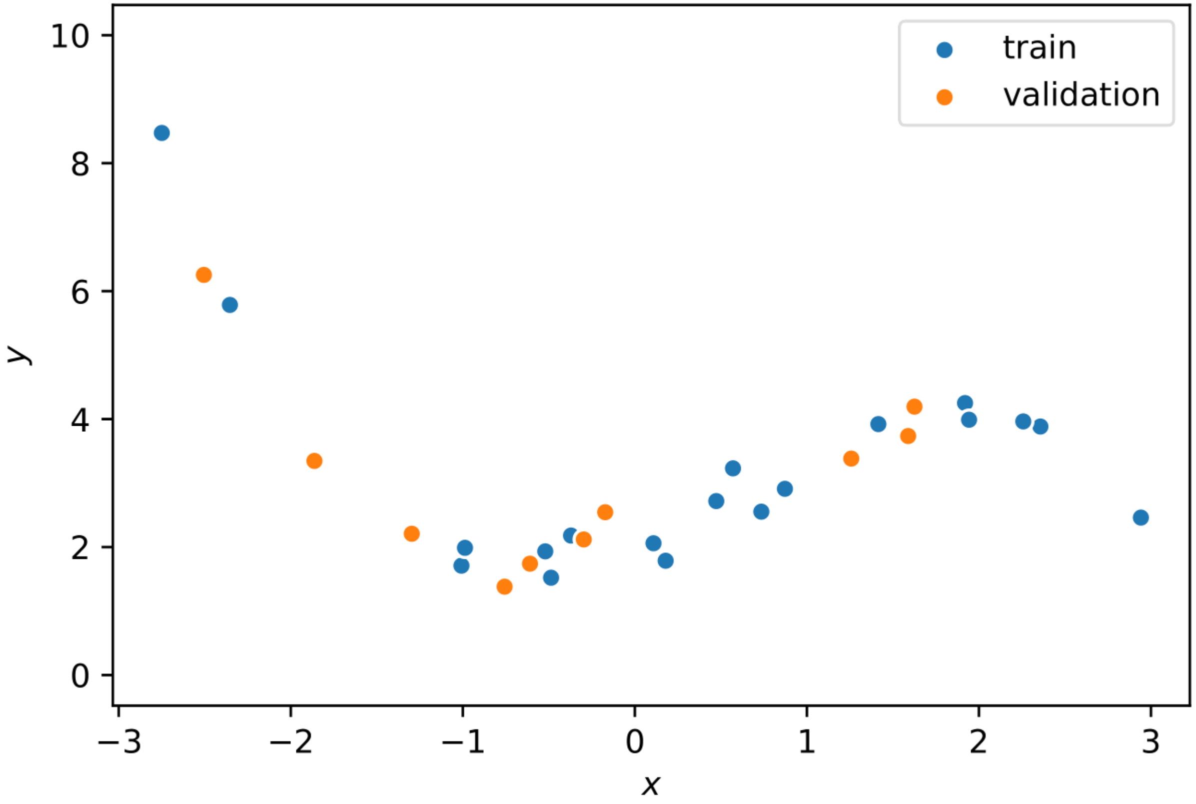


When reporting training loss as above, we don't include the regularization term so that it is comparable to the validation loss (where the regularization term is never included).

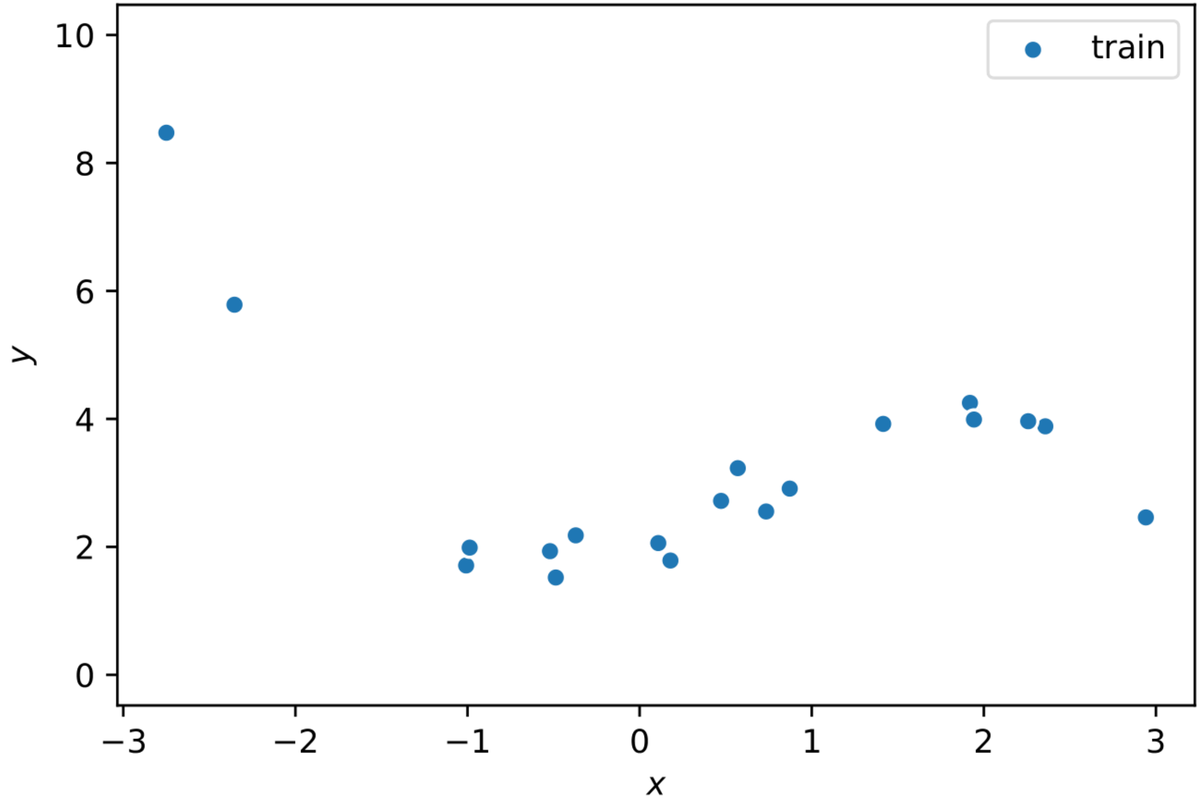
Which model should we use on some data?



Which model should we use on some data?

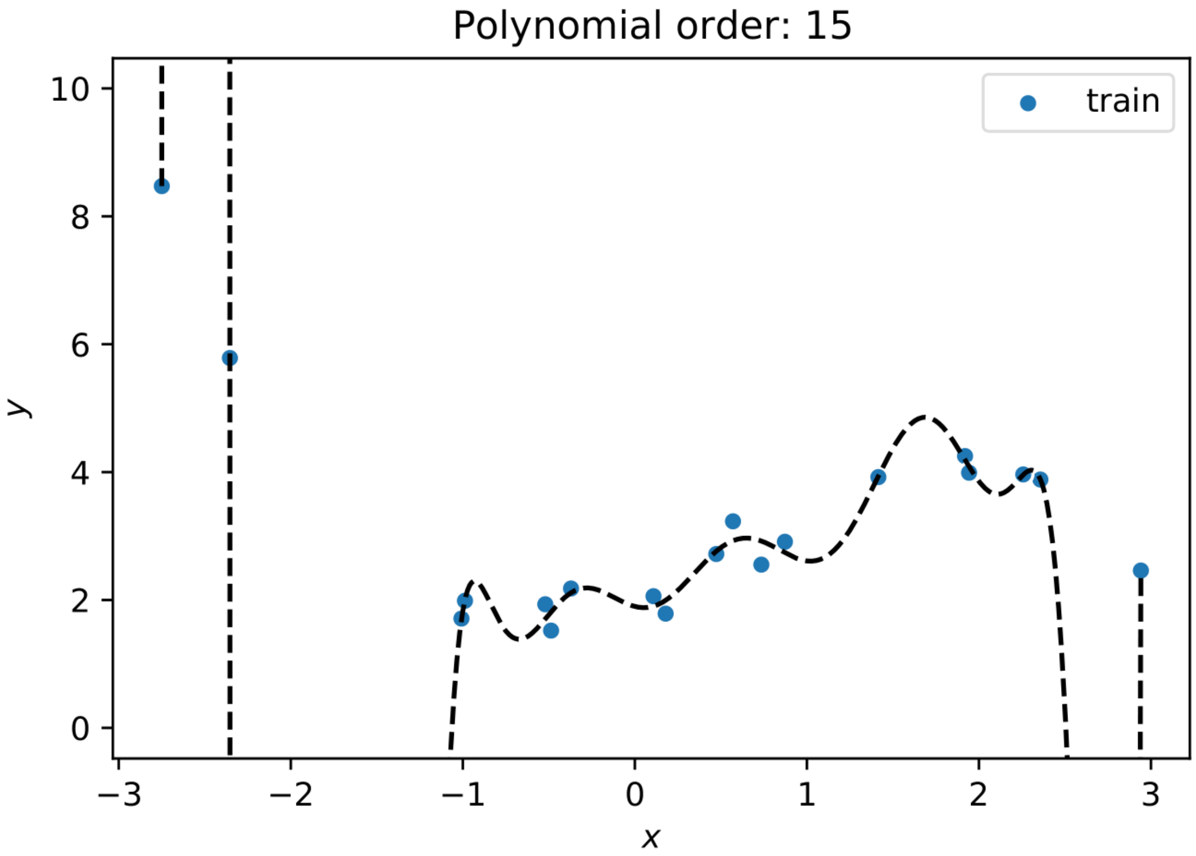


Which model should we use on some data?



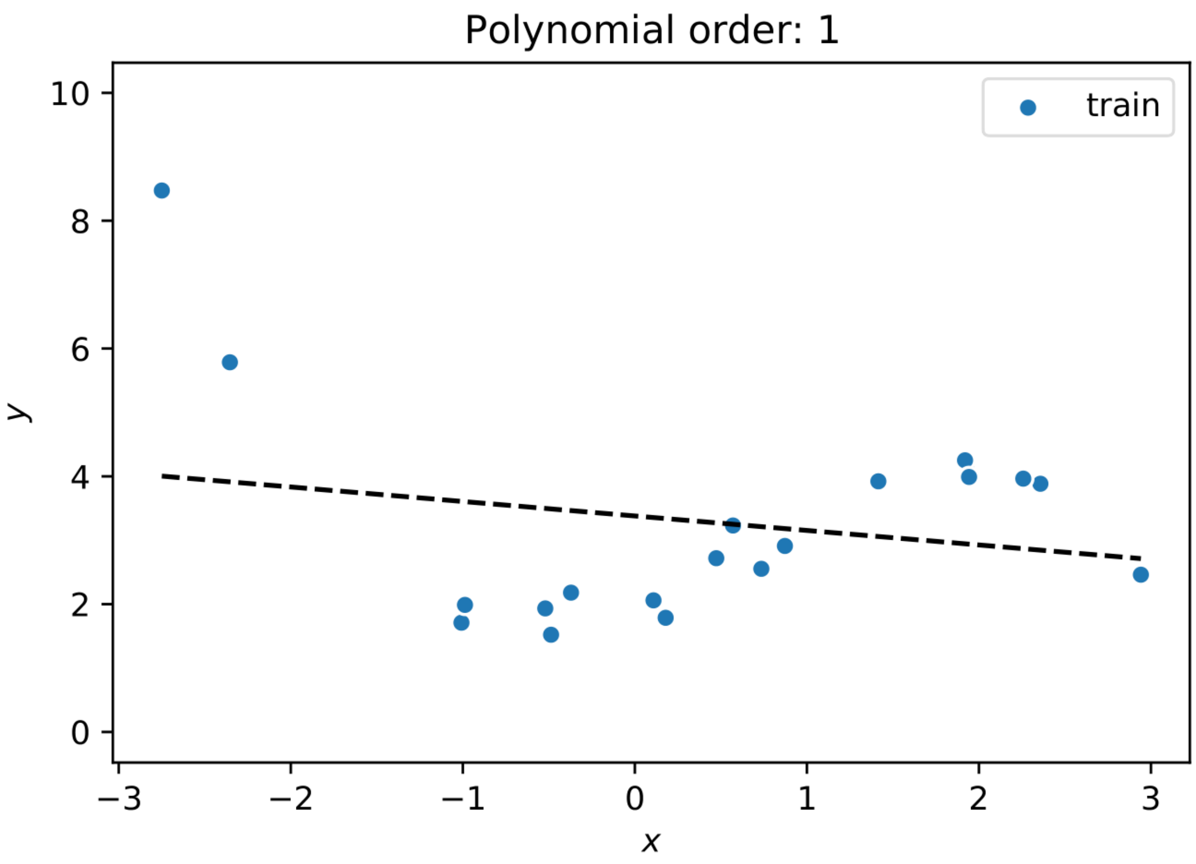
Let's fit some models on the training data

Which model should we use on some data?



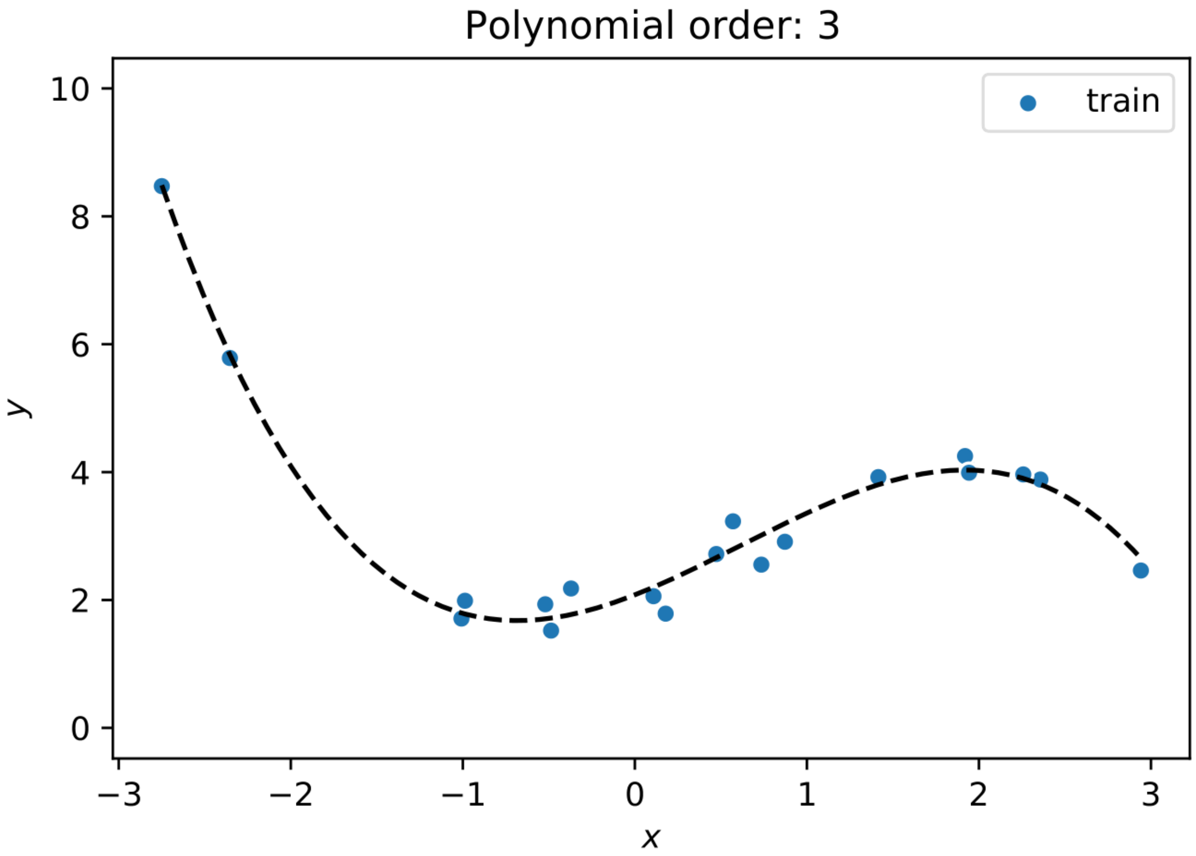
$$\mathrm{MSE}_{\mathrm{train}} = 0.0243$$

Which model should we use on some data?



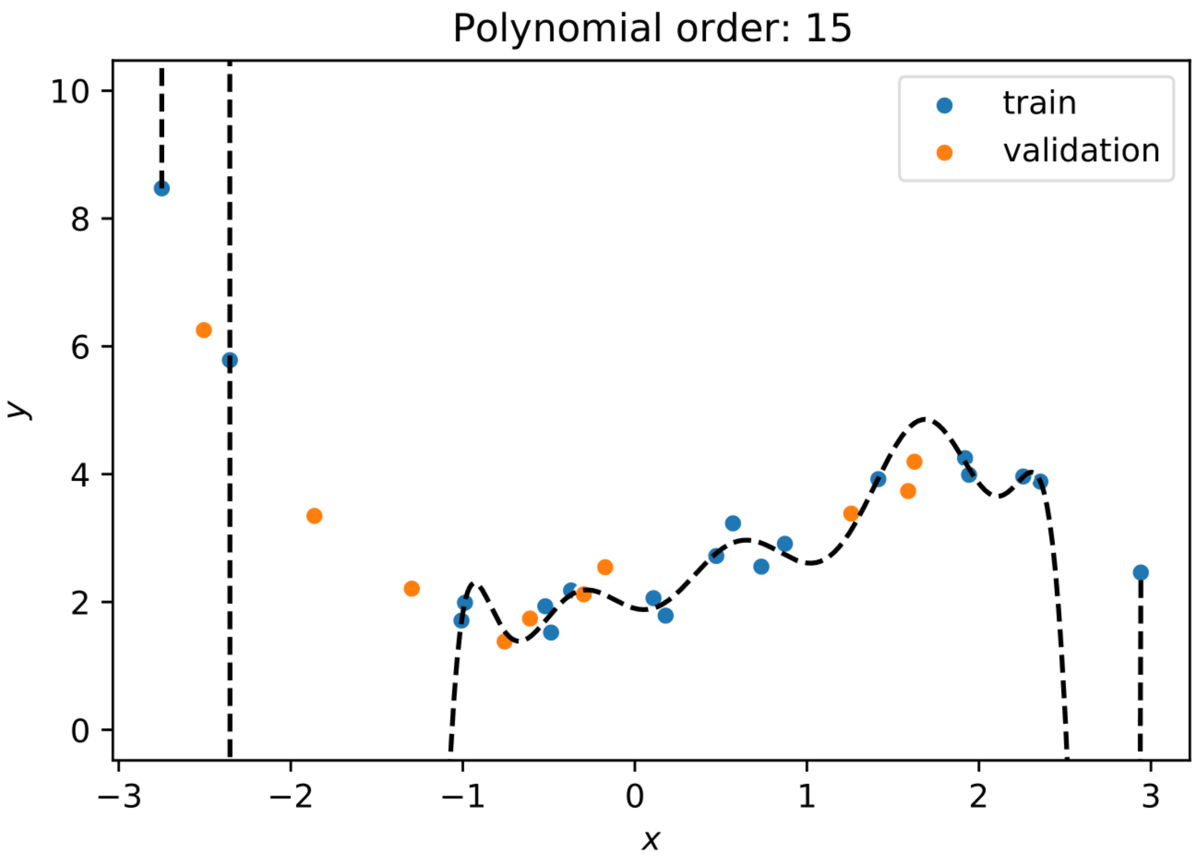
$\mathrm{MSE}_{\mathrm{train}} = 2.5005$

Which model should we use on some data?



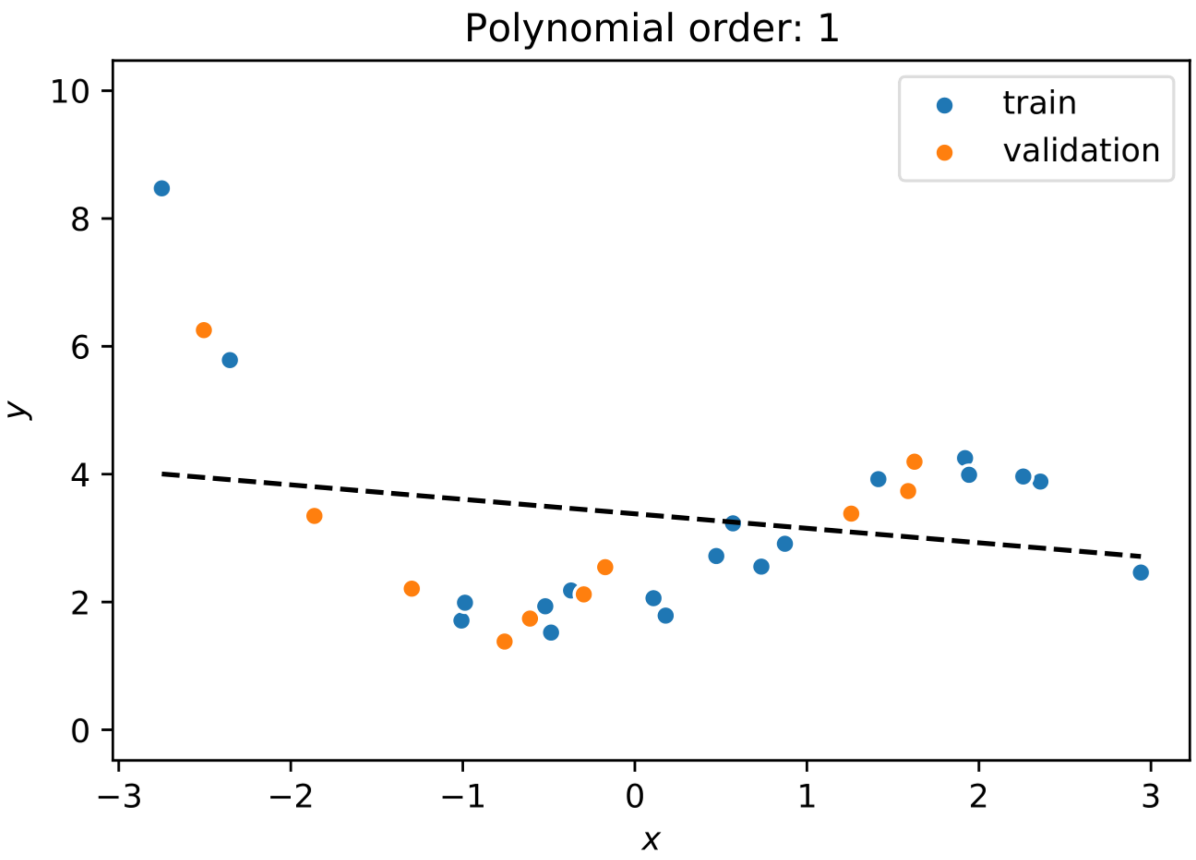
$$\mathrm{MSE}_{\mathrm{train}} = 0.0600$$

Which model should we use on some data?



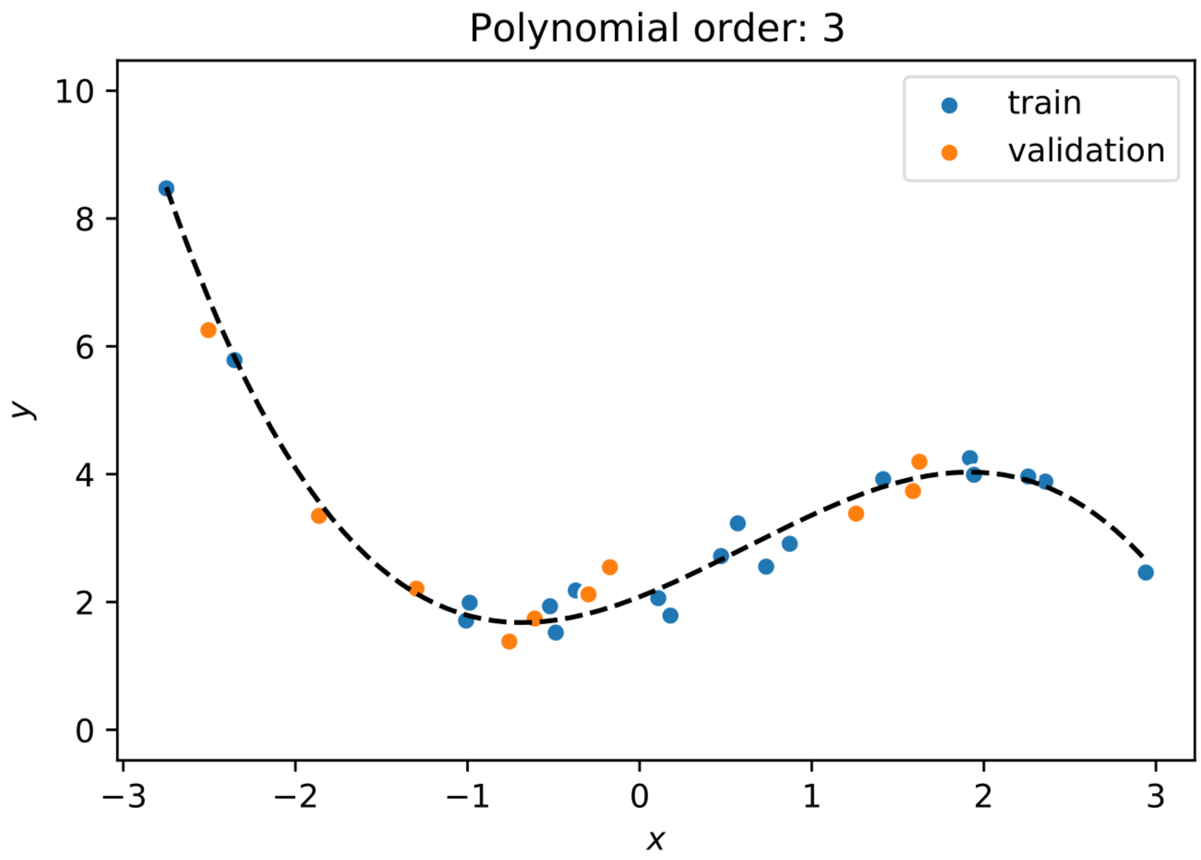
$\mathrm{MSE}_{\mathrm{train}} = 0.0243$; $\mathrm{MSE}_{\mathrm{val}} = 456386.9249$ (overfitting, “high variance”)

Which model should we use on some data?



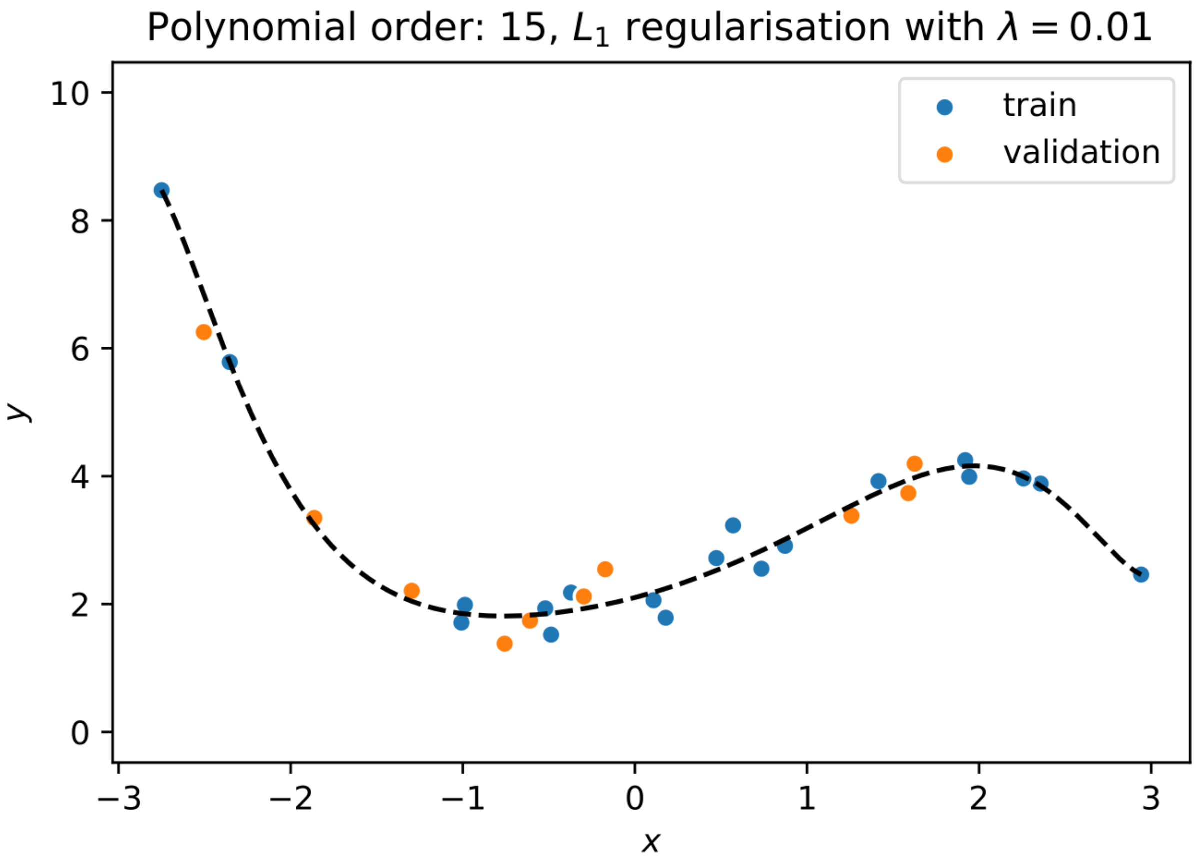
$\mathrm{MSE}_{\mathrm{train}} = 2.5005$; $\mathrm{MSE}_{\mathrm{val}} = 2.0069$ (underfitting, "high bias")

Which model should we use on some data?



$\mathrm{MSE}_{\mathrm{train}} = 0.0600$; $\mathrm{MSE}_{\mathrm{val}} = 0.1056$ (“just right”)

Which model should we use on some data?



$\mathrm{MSE}_{\mathrm{train}} = 0.0508$; $\mathrm{MSE}_{\mathrm{val}} = 0.1039$

Training on test data is one of the worst mistakes you can make in machine learning