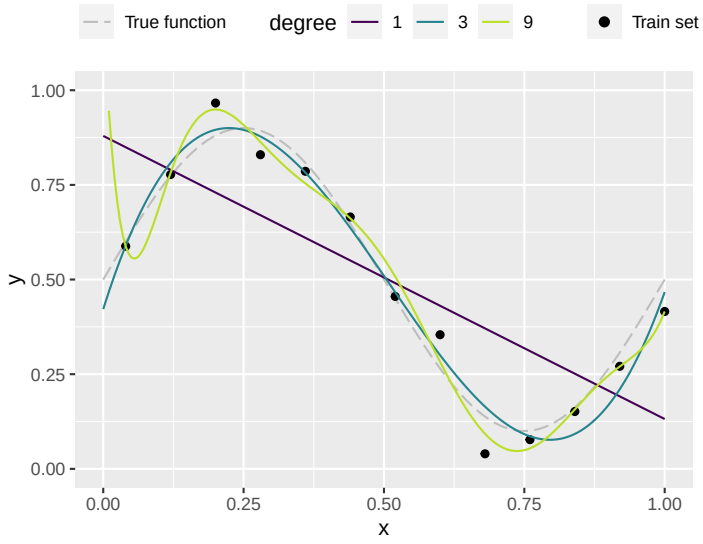
Introduction to Machine Learning

Evaluation: Train Error



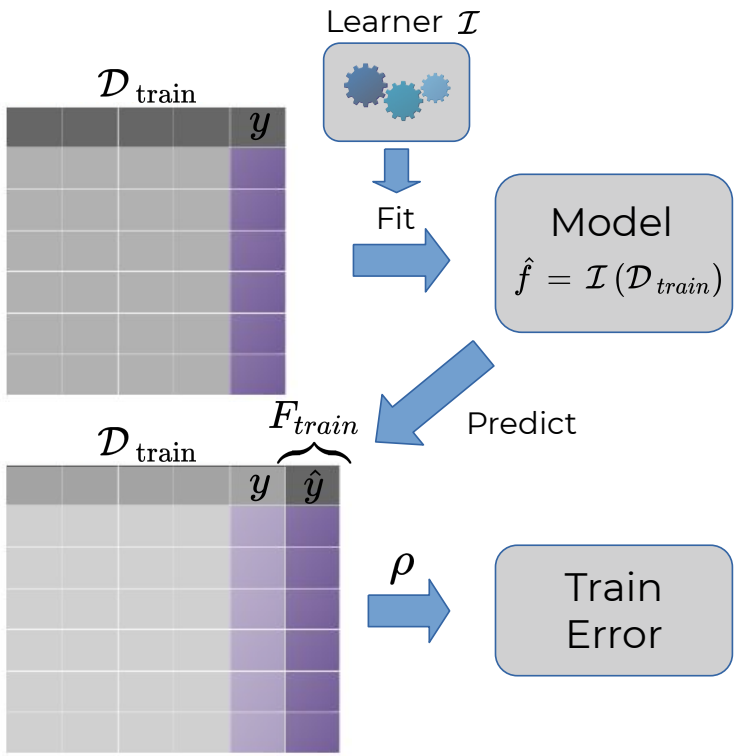
Learning goals

- Understand the definition of training error
- Understand why training error is no reliable estimator of future performance

TRAINING ERROR

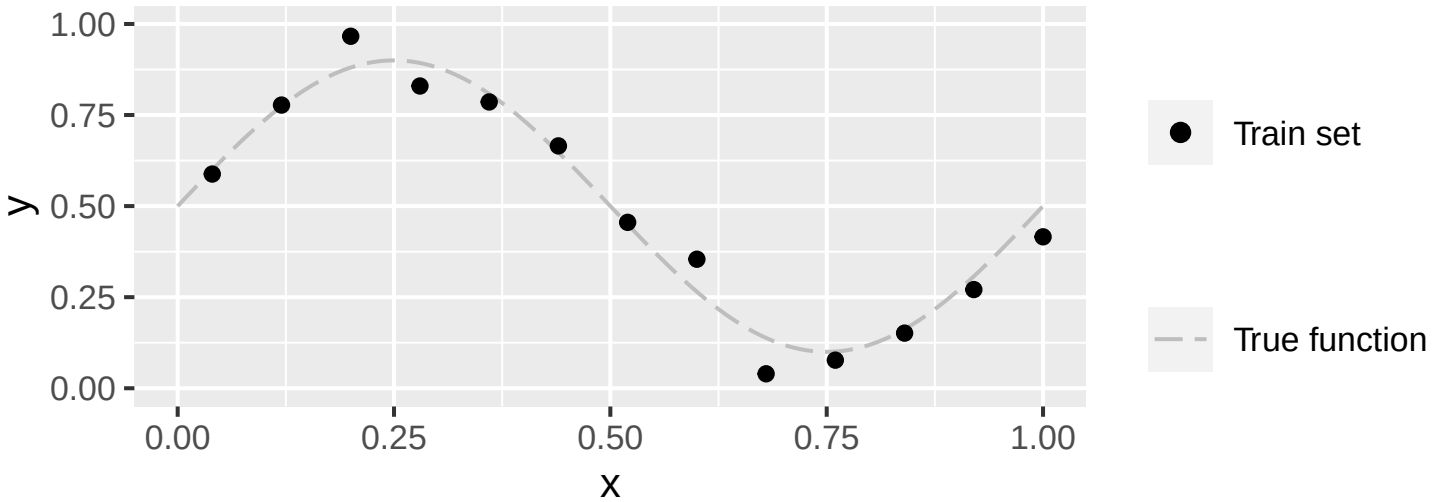
The **training error** (also called apparent error or resubstitution error) is estimated by averaging the errors over the training observations that have been used to fit the model:

$$\rho(\mathbf{y}_{\text{train}}, F_{\text{train}}) \text{ where } F_{\text{train}} = \left[\hat{f}_{\mathcal{D}_{\text{train}}}(\mathbf{x}^{(1)}_{\text{train}})^\top \quad \dots \quad \hat{f}_{\mathcal{D}_{\text{train}}}(\mathbf{x}^{(m)}_{\text{train}})^\top\right]^\top$$



EXAMPLE: POLYNOMIAL REGRESSION

Sample data from sinusoidal function $0.5 + 0.4 \cdot \sin(2\pi x) + \epsilon$ with measurement error ϵ:

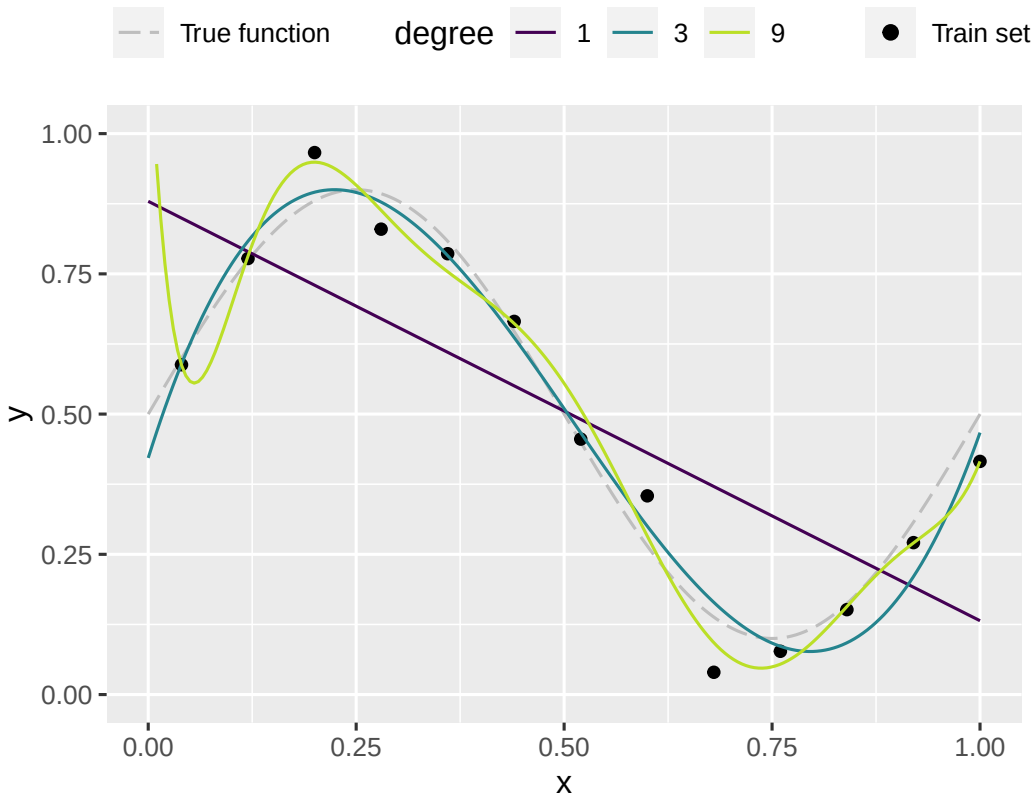


Now assume the data-generating process to be unknown as usual. Try to approximate the data with a d^{th}-degree polynomial:

$$f(\mathbf{x} \mid \boldsymbol{\theta}) = \theta_0 + \theta_1 \mathbf{x} + \cdots + \theta_d \mathbf{x}^d = \sum_{j=0}^{d} \theta_j \mathbf{x}^j.$$

EXAMPLE: POLYNOMIAL REGRESSION

Different polynomial orders give rise to models of varying **complexity**.
$\rightarrow$ How to choose d?



- $d = 1$: MSE = 0.036: clearly underfitting
- $d = 3$: MSE = 0.003: pretty OK
- $d = 9$: MSE = 0.001: clearly overfitting

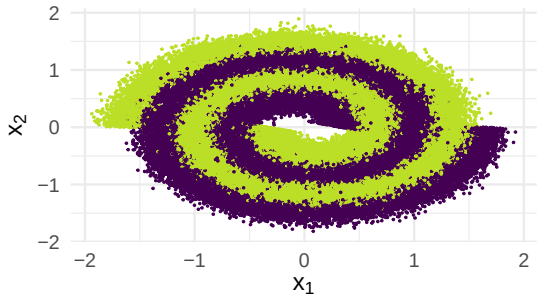
$\rightarrow$ Choosing d based on minimal training error seems to be a bad idea.

TRAINING ERROR PROBLEMS

- The training error is an unreliable and **optimistically biased** estimator of future performance.
  $\rightarrow$ Extreme example: training error of 1-NN is always 0 as each observation is its own NN at test time.
- We are interested in modeling the inherent data structure, not in fitting every peculiarity or noise in the training data.
- **Goodness-of-fit** measures like (classic) R^2, likelihood, AIC, BIC, deviance are all based on the training error.
- For models of restricted capacity, and given enough data, the training error may provide reliable information.
  - E.g., for a linear model with 5 features and 10^6 training points.
  - **But:** it is impossible to know when the training error starts to become unreliable.

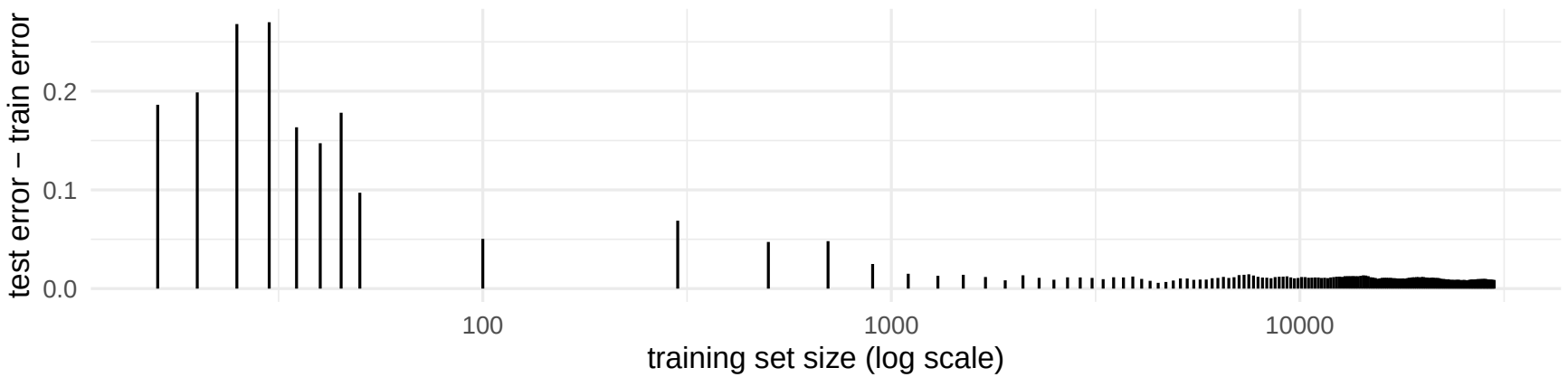
TRAINING ERROR AND GENERALIZATION

We simulate 100,000 points from the `spirals` data with $sd = 0.1$ and train a k-NN classifier with k set to 15.



For a sufficient amount of training data the training error should yield a fairly good indication of the true generalization ability.

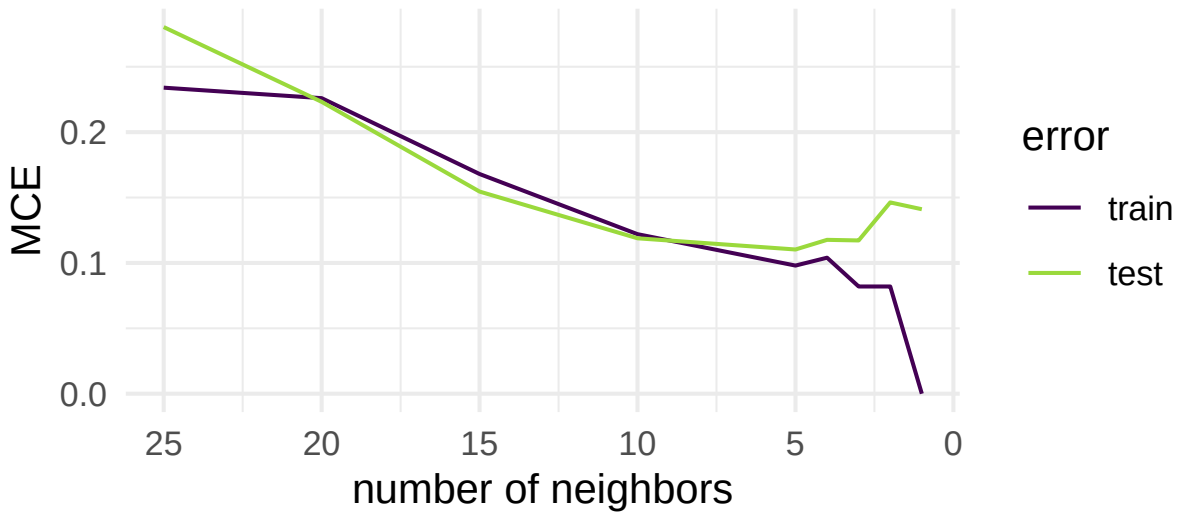
We set aside a very large test set of 70,000 observations to gauge the true generalization error as well as possible. Steadily increasing the training set size from 16 to 30,000 data points, we observe how the gap between training and test error narrows and bottoms out close to 0:



TRAINING ERROR AND GENERALIZATION

The example suggests that, for a large amount of training data and a rather simple learner, the training error can hold quite reliable information about the learner's ability to generalize.

But: assuming a more realistic scenario with a total of 500 training observations (while the test set remains the same) and allowing for more complexity, the picture changes:



The low training error for small k is deceptive – as the model becomes more and more local, test error starts creeping up as the learner fits the training data very closely at the expense of generalization.