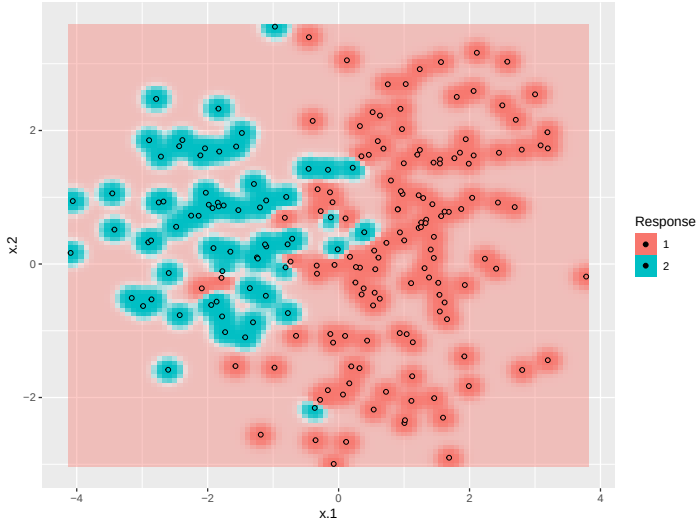
Introduction to Machine Learning

Evaluation: Overfitting

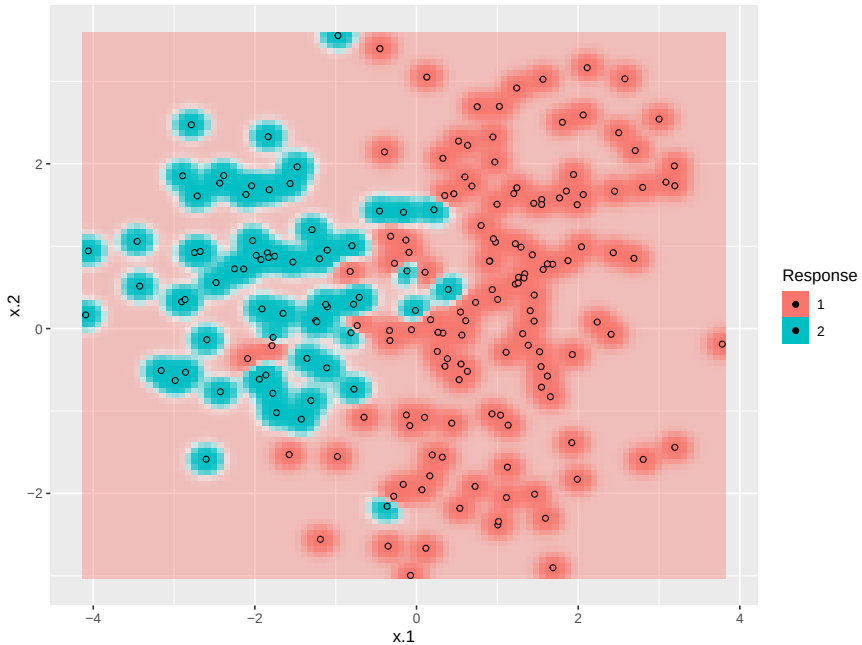


Learning goals

- Understand what overfitting is and why it is a problem
- Understand how to avoid overfitting

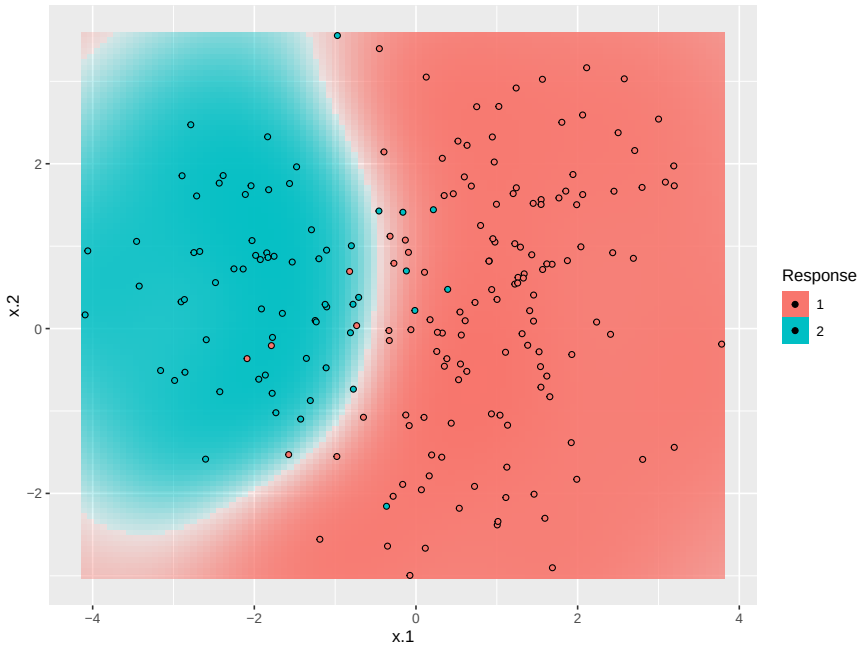
OVERFITTING

Overfitting learner



Better training set performance (seen examples)

Non-overfitting learner

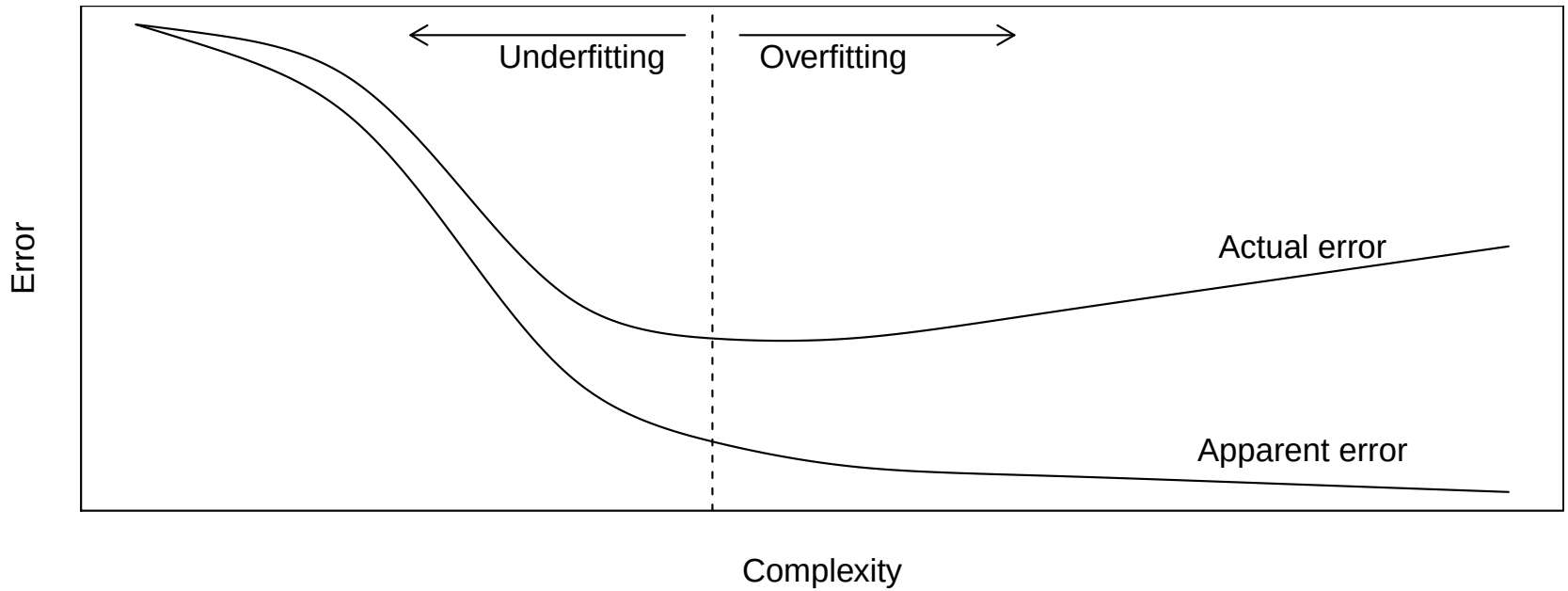


Better test set performance (unseen examples)

OVERFITTING

- Happens when algorithm models patterns beyond the data-generating process, e.g., noise or artefacts in the training data
- Reason: too many hypotheses and not enough data to tell them apart
- Less in bigger data sets
- If hypothesis space is not constrained, there may never be enough data
- Many learners have a parameter that allows constraining (*regularization*)
- Check for overfitting by validating on a new unseen test data set

TRADE-OFF BETWEEN GENERALIZATION ERROR AND COMPLEXITY



$\Rightarrow$ Optimization regarding model complexity is desirable:
Find the right amount of complexity for the given amount of data where generalization error becomes minimal.