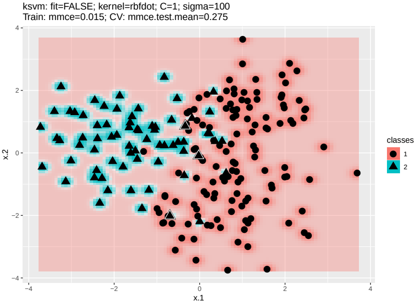
Introduction to Machine Learning

Evaluation: Overfitting

compstat-lmu.github.io/lecture_i2ml

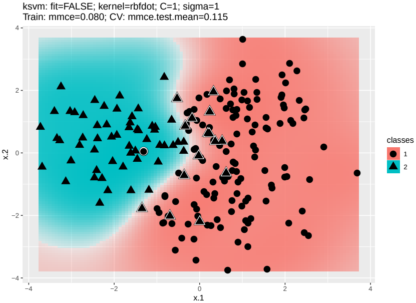
OVERFITTING

Overfitting learner



Better training set performance (seen examples)

Non-overfitting learner

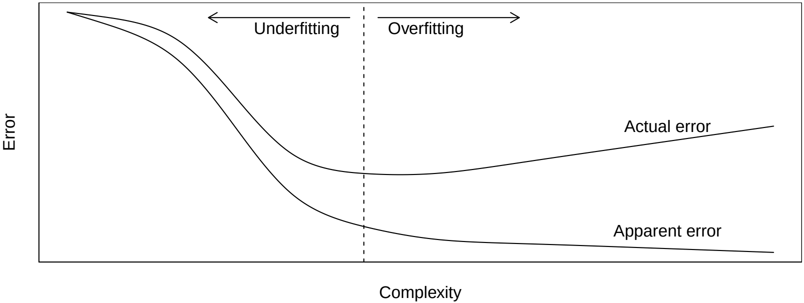


Better test set performance (unseen examples)

OVERFITTING

- Happens when algorithm models patterns beyond the data generating porcess, 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



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