LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Introduction to Machine Learning

Overfitting

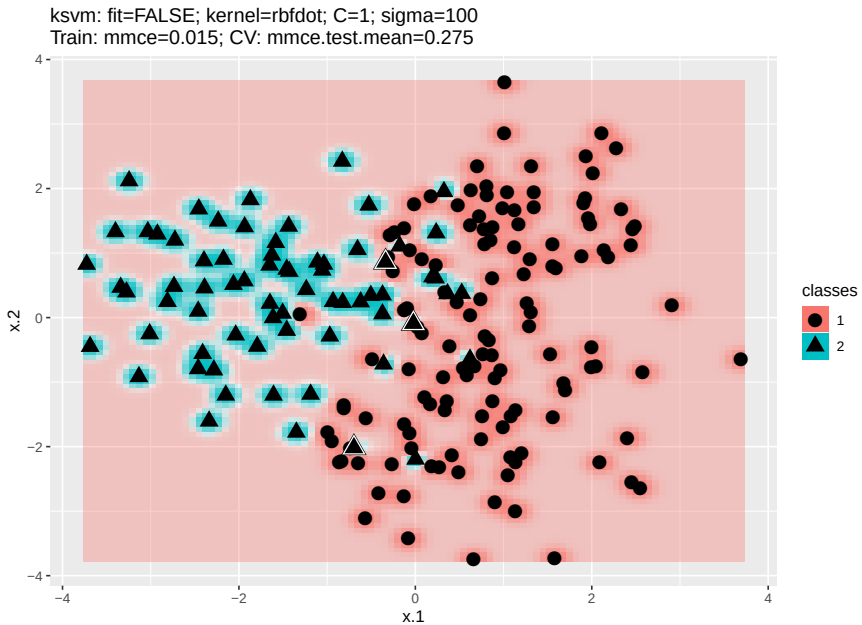
Department of Statistics – LMU Munich

OVERFITTING

- Overfitting is a well-known problem in ML for non-linear, powerful learning algorithms
- It happens when your algorithm starts modelling patterns in the data that are not actually true in the real world, e.g., noise or artefacts in the training data
- Happens when you have too many hypotheses and not enough data to tell them apart
- The more data, the more "bad" hypotheses are eliminated
- If the hypothesis space is not constrained, there may never be enough data
- There is often a parameter that allows you to constrain (*regularize*) the learner
- In this unit we will only give a very basic definition, and not really talk about measures against overfitting (see regularization!)

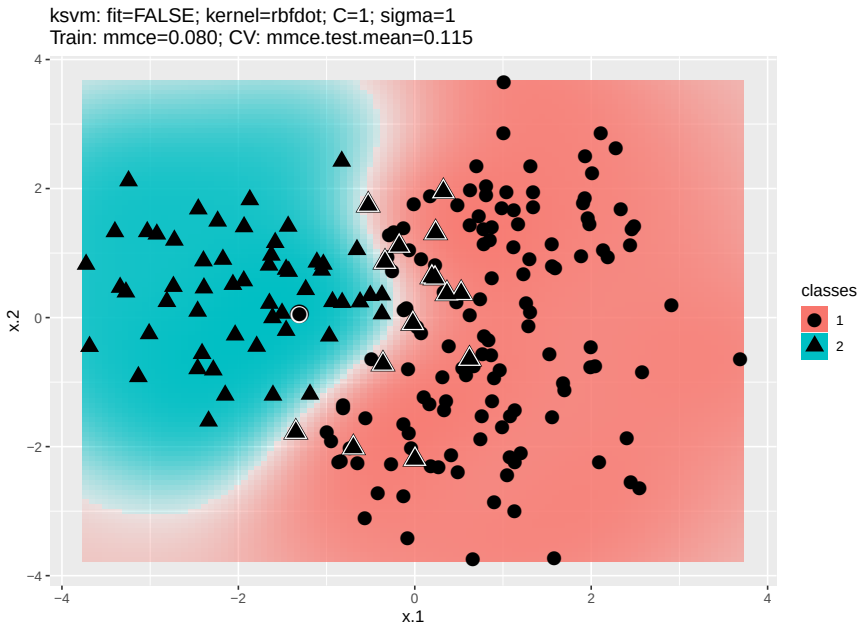
OVERFITTING

Overfitting learner



Better training set performance (seen examples)

Non-overfitting learner



Better test set performance (unseen examples)

OVERFITTING AND NOISE

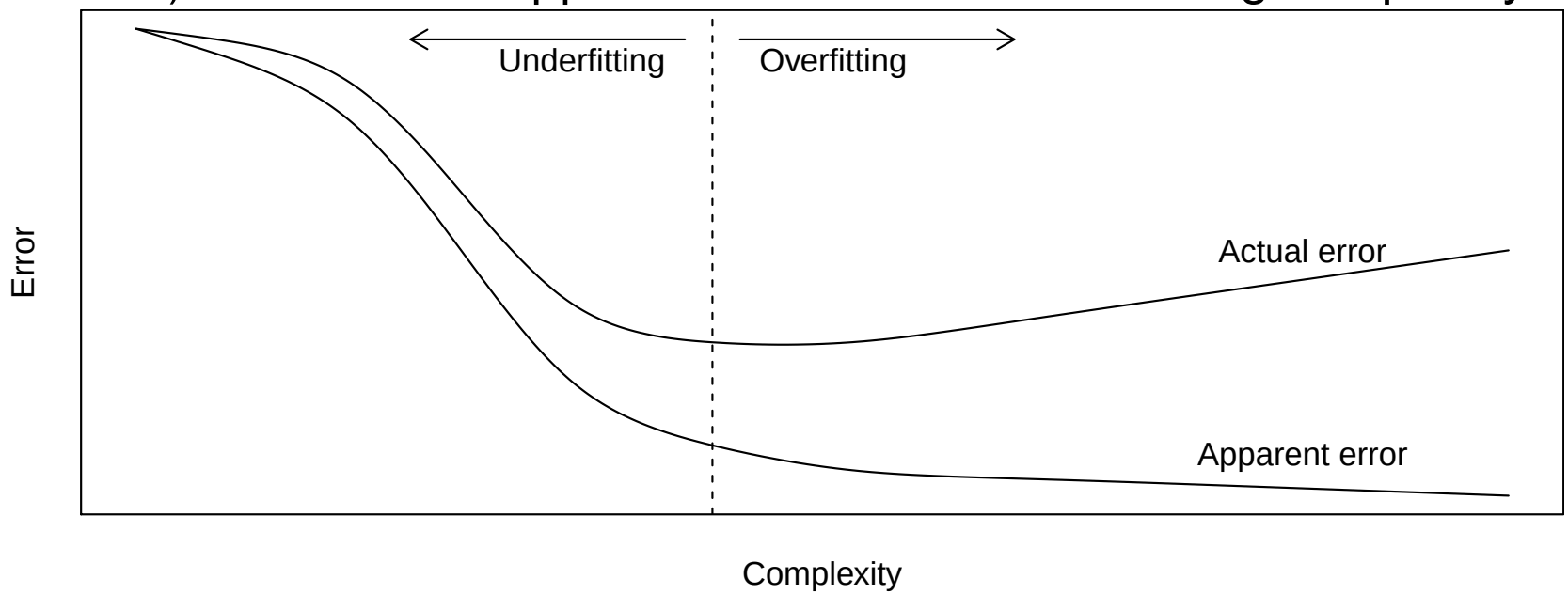
- Overfitting is seriously exacerbated by *noise* (errors in the training data)
- An unconstrained learner will start to model that noise
- It can also arise when relevant features are missing in the data
- In general it's better to make some mistakes on training data ("ignore some observations") than trying to get all correct

AVOIDING OVERFITTING

- You should never believe your model until you've *verified it on data that the learner didn't see*
- Scientific method applied to machine learning: model must make new predictions that can be experimentally verified
- Use less complex models
- Get more, or better data
- Some learner can do "early stopping" before perfectly fitting (i.e., overfitting) the training data
- Use regularization

TRADE-OFF BETWEEN GENERALIZATION ERROR AND COMPLEXITY

Apparent error (on the training data) and real error (prediction error on new data) evolve in the opposite direction with increasing complexity:



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