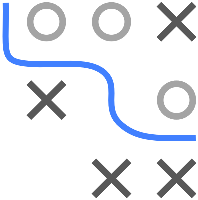
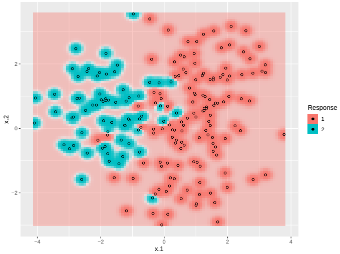
# Introduction to Machine Learning



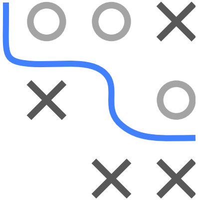
# Evaluation
# Overfitting and Underfitting




### Learning goals

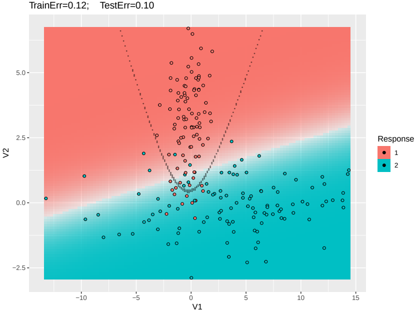
- Understand definitions of overfitting and underfitting

# UNDERFITTING

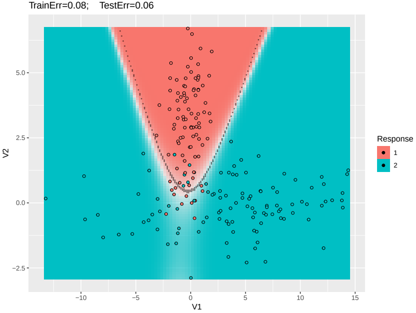


- Occurs if model does not reflect true shape of underlying function
- Hence, predictions will be less good as they could be
- High train error and high test error
- Hard to detect, as we don't know what the Bayes error is for a task

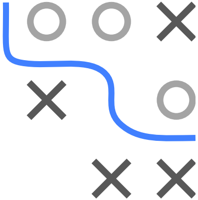
Underfitted model




Appropriate model

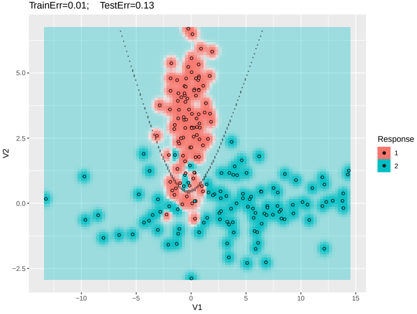


# OVERFITTING

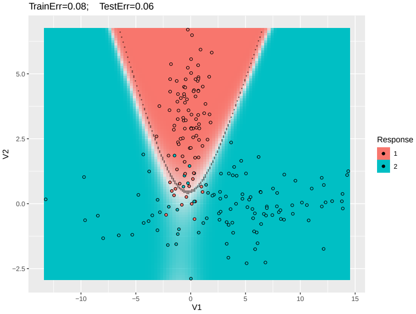


- Overfitting occurs when the model reflects noise or artifacts in training data, which do not generalize
- Small train error, at cost of test high error
- Hence, predictions of overfitting models cannot be trusted - but proper ML evaluation workflows should make it visible

Overfitted model

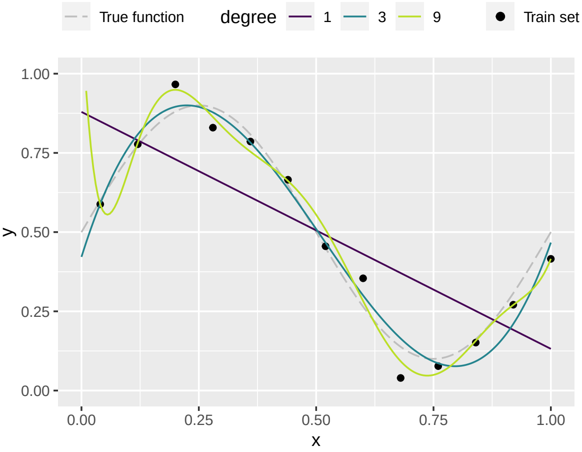



Appropriate model

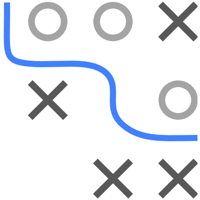


# UNDER- AND OVERFITTING IN REGRESSION

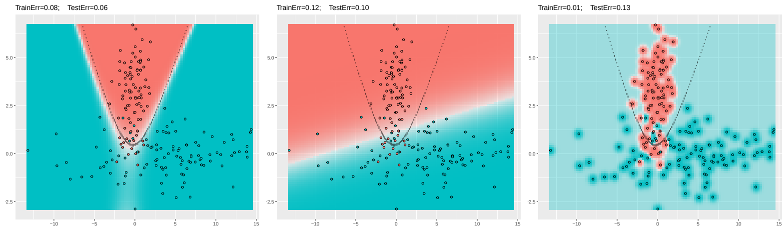
- Poly-Regression, on data from sinusoidal function
- LM underfits, high-d overfits



# MATHEMATICAL DEFINITIONS

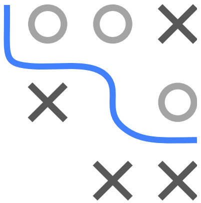


- Nearly no reference does that, here is one approach
- Underfitting $UF(\hat{f}, L) := \mathrm{GE}\left(\hat{f}, L\right) - \mathrm{GE}(f^*, L)$
  Diff in GE between $\hat{f}$ and the Bayes optimal model
- Overfitting $OF(\hat{f}, L) := \mathrm{GE}\left(\hat{f}, L\right) - \mathcal{R}_{\text{emp}}(\hat{f}, L)$
  Diff between (theoretical) GE and training error




NB: Now, RHS is both UF and OF, let's say OF has "prio".

# OVERFITTING TRADE-OFFS



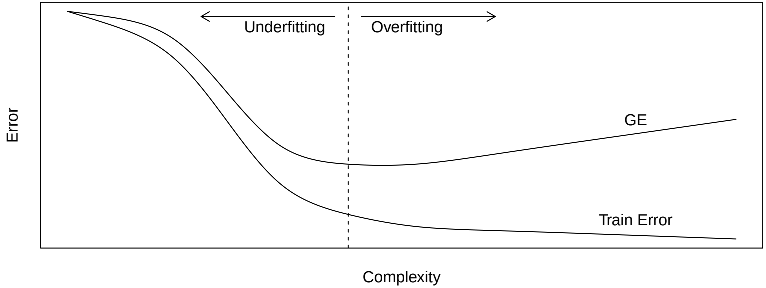
The potential for overfitting is influenced by:

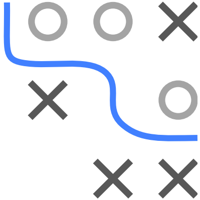
- Complexity of hypothesis space
- Amount of training data
- Dimensionality of feature space
- Irreducible noise

Implications:

- The larger / more complex is $\mathcal{H}$, the more data we need to tell candidate models apart
- The less data we have, the more we need to stick with "constrained" $\mathcal{H}$
- OF can happen for LMs too: If feature space is very high-dim
- Tightly connected to the bias-var-noise decomposition of GE of a learner ($\rightarrow$ which we study elsewhere).

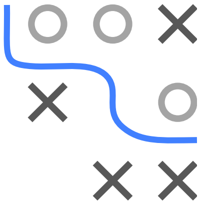
# COMPLEXITY VS GE



- Common U-shape of GE if complexity or train-rounds go up.
- Optimal level of complexity:
  Simplest model for which GE is not significantly outperformed
- We could also call "Point of OF" the point where GE goes up.

# AVOIDING OVERFITTING



- Use more or better data – not always possible, but maybe can augment data, e.g., for images
- Constrain $\mathcal{H}$ directly by using less complex model classes
- Many learners come with HPs that can constrain complexity
- Use "early-stopping"
- Occam's razor in model selection: If GE not strongly reduced for more complex class, use the simpler model.

All of the above are methods of regularization, which we study in a dedicated chapter.