The Machine Learning Basics

Chapter 5

*“A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P**, if its performance at tasks in T, as measured by P, improves with experience E.”*

- Mitchell (1997)

Generalization

- Generalization error
- In the linear regression, we train the model by minimizing the training error,

$$\frac{1}{m^{(\text{train})}}||\boldsymbol{X}^{(\text{train})}\boldsymbol{w} - \boldsymbol{y}^{(\text{train})}||_2^2,$$

but we actually care about the test error.

$$\frac{1}{m^{(\text{test})}}||\boldsymbol{X}^{(\text{test})}\boldsymbol{w} - \boldsymbol{y}^{(\text{test})}||_2^2.$$

Training error and test error

- i.i.d assumptions
  - Independant
  - Identically distributed
- The probabilistic framework and the i.i.d. assumptions allow us to study the relationship between training error and test error.

“The factors determining how well a machine learning algorithm will perform are its ability to: ***Make the training error small*** *and* ***Make the gap between training and test error small.***”

Underfitting and Overfitting

Underfitting occurs when the model is not able to obtain a sufficiently low error on the training set.

Overfitting occurs when the gap between the training error and test error is too large.

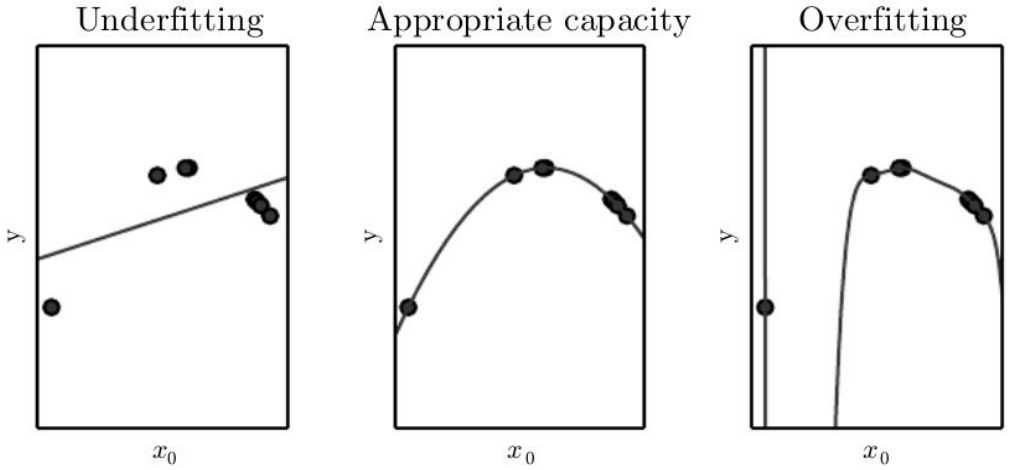
Capacity

- Choosing the hypothesis space
- Occam’s razor principle:
  - Among competing hypotheses that explain known observations equally well, one should choose the “simplest” one.
- Gap between training error and generalization error grows as the model capacity grows but shrinks as the number of training examples increases.

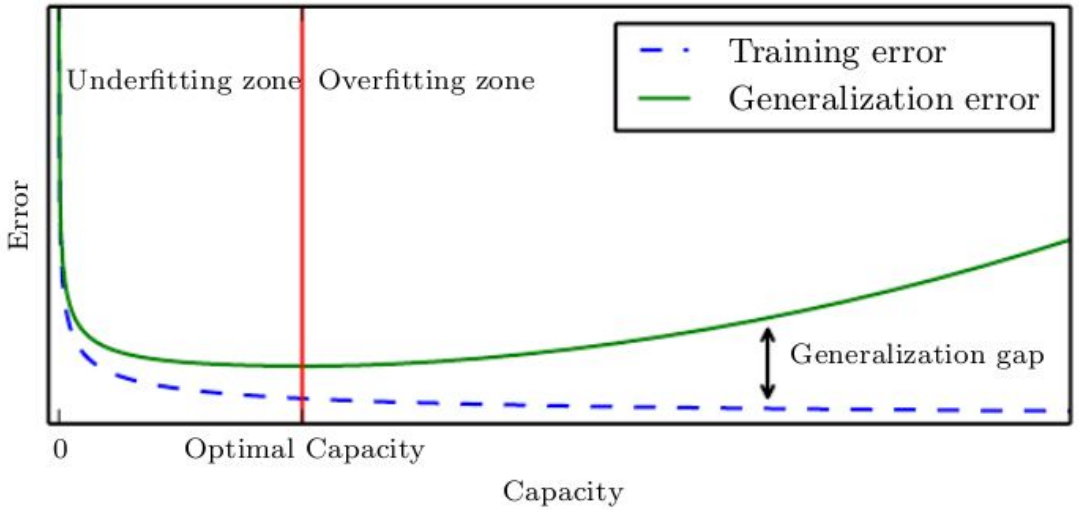
Left: A linear function cannot capture the curvature that is present in the data.

Center: A quadratic function fit to the data generalizes well to unseen points.

Right: A higher degree polynomial fit to the data suffers from overfitting.



As we increase capacity, training error decreases, but the gap between training and generalization error increases.



The No Free Lunch Theorem

“Averaged over all possible data generating distributions, every classification algorithm has the same error rate when classifying previously unobserved points. In other words, in some sense, no machine learning algorithm is universally any better than any other.”

-Walpert, 1996

Regularization

- Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error.
- Also minimizing J(w) that expresses a preference for the weights to have smaller squared L2 norm,

$$J(\boldsymbol{w}) = \mathrm{MSE}_{\mathrm{train}} + \lambda \boldsymbol{w}^{\top} \boldsymbol{w},$$

Hyperparameters

- Settings to control the behavior of the learning algorithm
- Validation Sets

Thank you