Section 2E. Model Quality

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

Assessing Model Quality

We should differentiate between two types of data:

- ***Training dataset*** $\mathcal{D}_{\text{Tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$: Data available while estimating the unknown parameters θ of your parametric model, $\widehat{f}(\mathbf{x};\theta)$. The ***training MSE*** is defined as

$$\text{MSE}_{\text{Tr}} = \frac{1}{N} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\textbf{Tr}}} \left(y_i - \widehat{f}(\mathbf{x}_i;\theta) \right)^2$$

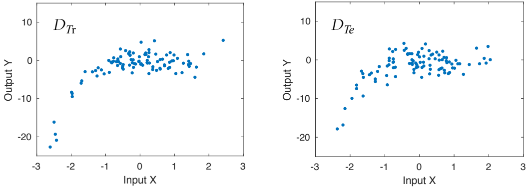
- ***Testing dataset*** $\mathcal{D}_{\text{Te}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{M}$: Data that your learning algorithm has *not* seen during training and can be used to estimate the performance of future predictions using the ***test MSE***

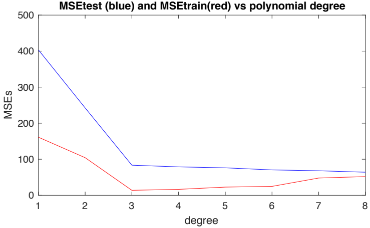
$$\text{MSE}_{\text{Te}} = \frac{1}{M} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\textbf{Te}}} \left(y_i - \widehat{f}(\mathbf{x}_i;\theta) \right)^2$$

The test MSE is a better reflection of the performance of your model.

Model Selection

Consider a training and a testing datasets, $\mathcal{D}_{\text{Tr}}$ and $\mathcal{D}_{\text{Te}}$ containing samples $(\mathbf{x}_i, y_i) \sim f_{X,Y}$

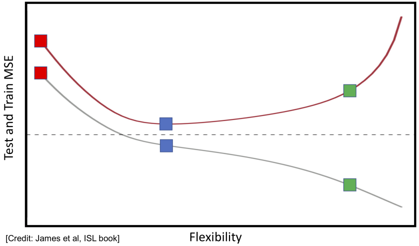




Model Selection (cont.)

Comments on test and train MSE (conceptual figure below):

- ***Train MSE*** (gray plot) decays monotonically as the flexibility increases.
- ***Test MSE*** (red plot) presents an optimal minimum value (blue square).
  - Below the optimal flexibility level, the model $\widehat{f}$ is not flexible enough to learn f faithfully (red square)
  - Above the optimal flexibility level, the model $\widehat{f}$ is too flexible and starts following the noise in our training data (gree square)



[Credit: James et al, ISL book]

Penn Engineering

Copyright 2020 University of Pennsylvania
No reproduction or distribution without permission.