Data-Driven Problem Solving in Mechanical Engineering

Polynomial Regression and Model's Capacity

Masoud Masoumi

ME 364 - Spring 2022

Department of Mechanical Engineering
The Cooper Union for the Advancement of Science and Art

March, 2022

Polynomial Regression as MLR

In polynomial regression, the relationship between the independent variable **x** and the dependent variable (target variable) **y** is modeled as an n^{th} degree polynomial in **x**.

Consider a polynomial as
$y = \omega_o + \omega_1 x + \omega_2 x^2 + \omega_3 x^3 + ... = \mathrm{w}^T \mathrm{x}$
with
$\mathrm{w}^T = [\omega_o, \omega_1, \omega_2, \omega_3, ...]$ and $\mathrm{x} = [1, x, x^2, x^3, ...]^T$

We create a few additional features: 1, x, x^2, ...

$$\begin{cases} x & = x_1 \\ x^2 & = x_2 \\ x^3 & = x_3 \\ ... & \end{cases}$$

Polynomial Regression as MLR

so the polynomial becomes a multiple linear regression.
$y = \omega_o + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3 + ...$

So if we take a specific column in our data set as feature x, here is what this transformation will do for an n^{th} degree polynomial

$$\begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{bmatrix} \longrightarrow \begin{bmatrix} [1 & x^{(1)} & x^{(1)^2} & ... & x^{(1)^n}] \\ [1 & x^{(2)} & x^{(2)^2} & ... & x^{(2)^n}] \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ [1 & x^{(m)} & x^{(m)^2} & ... & x^{(m)^n}] \end{bmatrix}$$

Fitting Problem

The central challenge in machine learning is that our algorithm must perform well on new, previously unseen inputs - not just those on which our model was trained. The ability to perform well on previously unobserved inputs is called **generalization**.

The factors determining how well a model will perform are its ability to:

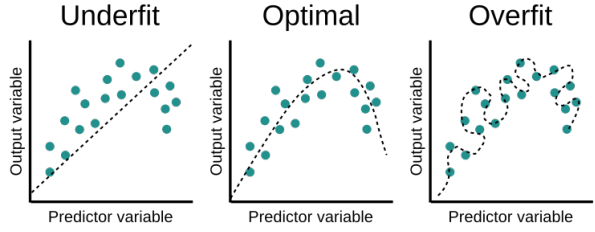
- Make the training error small (optimization).
- Make the gap between training and test error small (generalization).

These two factors correspond to the two central challenges in machine learning: **underfitting** and **overfitting**.

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

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

Fitting Problem



- An underfit model fails to significantly grasp the relationship between the input values and target variables. This may be the case when the model is too simple

- An overfit model has overly memorized the data set it has seen and is unable to generalize the learning to an unseen data set. That is why an overfit model results in very poor test accuracy. This may occur when the model is highly complex

Fitting Problem

We can control whether a model is more likely to overfit or underfit by altering its **capacity**.

Informally, a model's capacity is its ability to fit a wide variety of functions. Models with low capacity may struggle to fit the training set. Models with high capacity can overfit by memorizing properties of the training set that do not serve them well on the test set.