Generalization

Fatemeh Seyyedsalehi

References of the lecture are mentioned in the last slide

Where are we now?

- We will attempt to understand the generalization of the learning models, i.e. their performance on unseen data.

- Up to now, we have found the model $h_w(\boldsymbol{x})$, which minimizes a training cost function $J(\boldsymbol{w})$ over the training set, $\mathrm{D} = \{(\boldsymbol{x}^i, \boldsymbol{y}^i), i = 1, \dots, n\}$.
  - For example, the SSE cost function:

$$J(\boldsymbol{w}) = \sum_{i=1}^{n} \left(y^{(i)} - h_w(x^{(i)})\right)^2$$

Where are we now?

- However, it is not our main goal.
  - In fact, it is our approach towards the goal of learning a predictive model.

- The most important evaluation metric is the model performance on unseen test example, which is called the test error.

Where are we now?

- The key difference between the training set and test examples are that the test examples are unseen during the training procedure.

- Therefore, the test error is not necessarily close to the training error.

Where are we now?

- Overfitting: The model predicts accurately on the training dataset but doesn't generalize well to other test examples, that is, if the training error is small but the test error is large.

- Underfitting: The training error is relatively large, in this case, typically the test error is also relatively large.

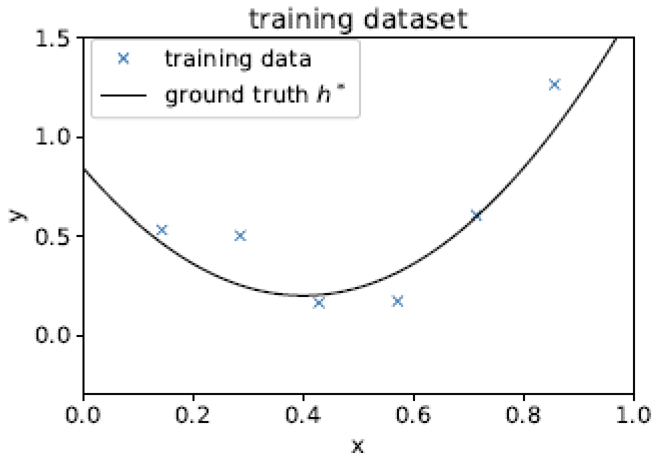
Bias-variance tradeoff
Example

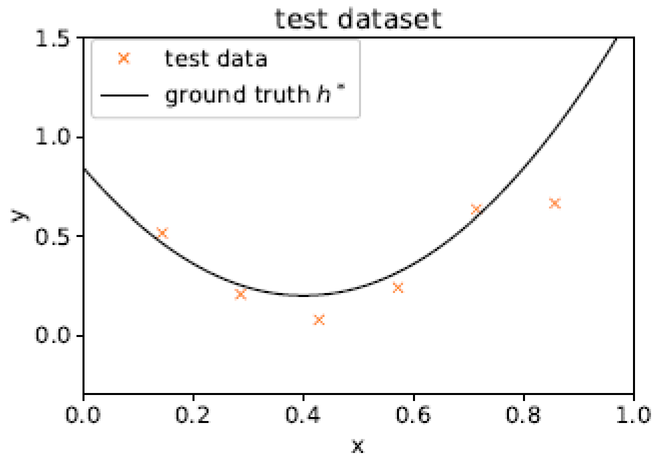
- As an example consider the following target function,

$$y = h_w^t(x) + \varepsilon$$

where $h_w^t(x)$ is a quadratic form and ε is the noise observation.

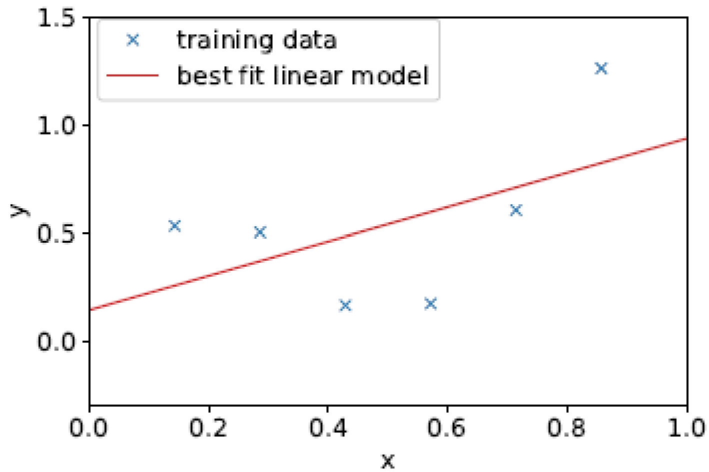
- It is impossible to predict the noise ε and our goal is to recover the function $h_w^t(x)$.

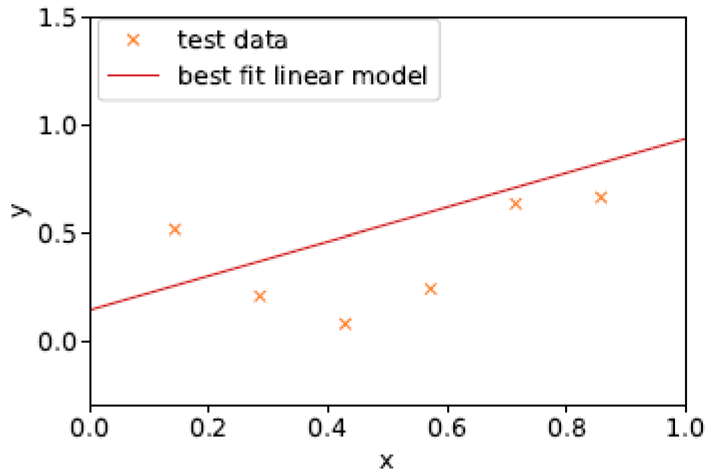




Bias-variance tradeoff Example

- The hypothesis space of “linear models”

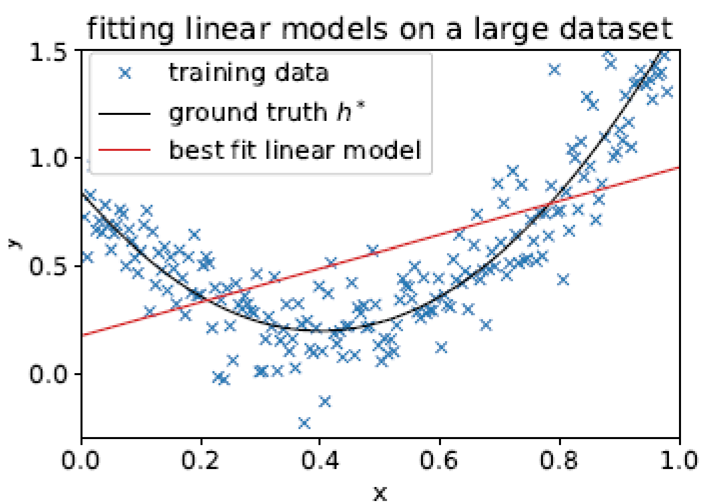


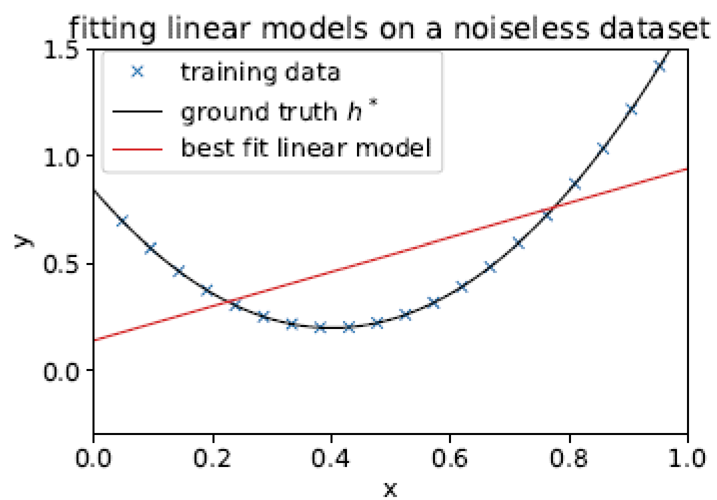


- The best linear model has large training and test errors
  - Underfitting

Bias-variance tradeoff
Example

- The hypothesis space of "linear models"
- Even with a very large amount of, or even infinite training examples, the best fitted linear model is still inaccurate and fails to capture the structure of the data.
  - This issue still occurs even without the presence of the noise





Bias-variance tradeoff
Example

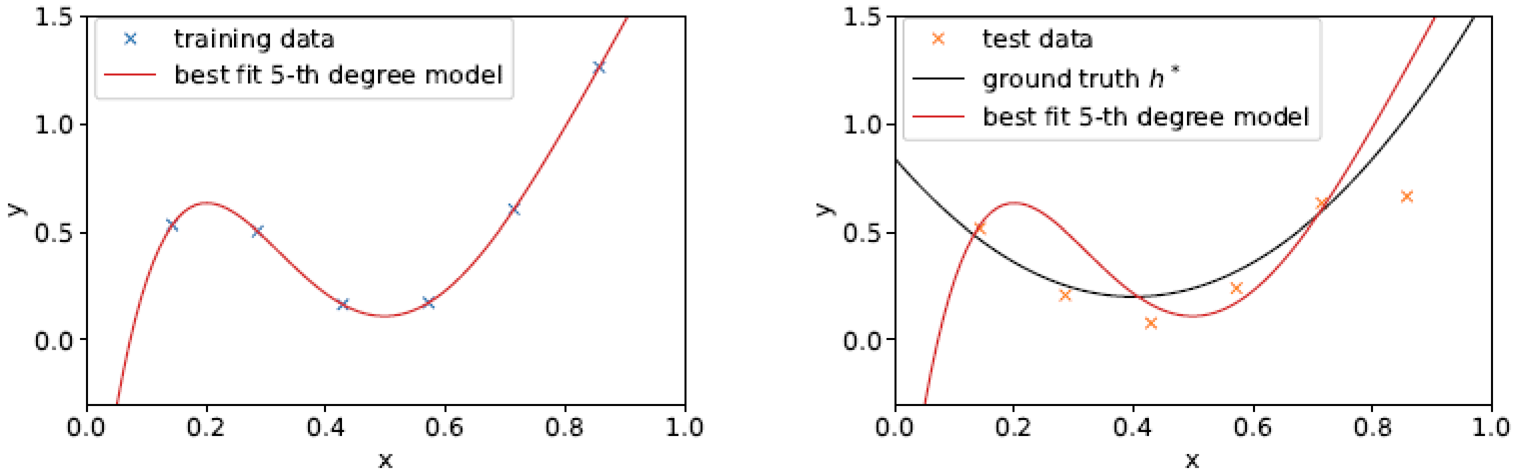
- The main problem is that linear models can not represent a quadratic function $h_w^t(x)$.

- Informally, we define the **bias** of a model to be the test error when we train the model with a very large (infinite) training dataset.

- In this example, the linear model suffers from large **bias**, and underfits, i.e. fails to capture the structure exhibited by the data.

Bias-variance tradeoff Example

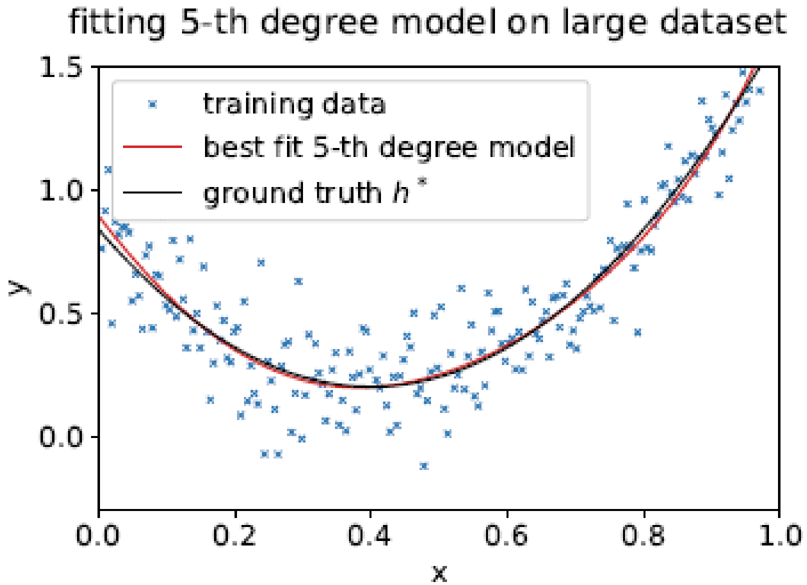
- The hypothesis space of “5th degree polynomial”



- The model learnt from the training set does not generalize well to test examples.
- Best 5th degree polynomial has zero training error, but still has a large test error
  - Overfitting

Bias-variance tradeoff
Example

- The hypothesis space of "5th degree polynomial"
- Fitting to an extremely large dataset, the resulting model would be close to a quadratic function and be accurate.
  - This is because the family of 5th degree polynomials contains all the quadratic functions and capable of capturing the structure of the data.

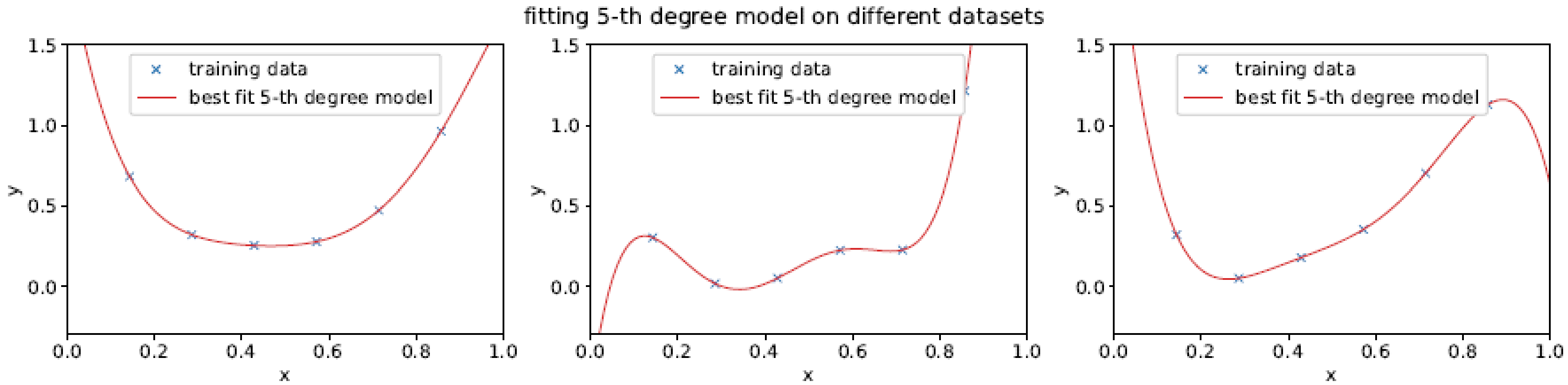


Bias-variance tradeoff
Example

- The failure of fitting 5th degree polynomials has different reason from failure of linear models.
  - The best 5th degree polynomial on a huge dataset nearly recovers the ground-truth
- Their failure can be captured by another component of the error called **variance.**

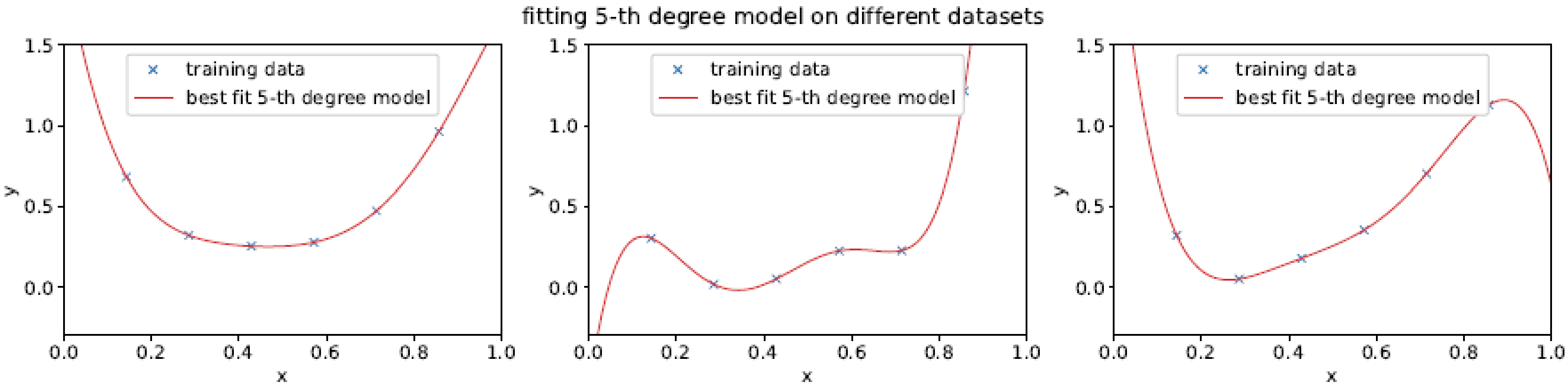
Bias-variance tradeoff
Example

- The best 5th degree models on three different datasets generated from the same distribution behave quite differently, suggesting the existence of a large **variance.**



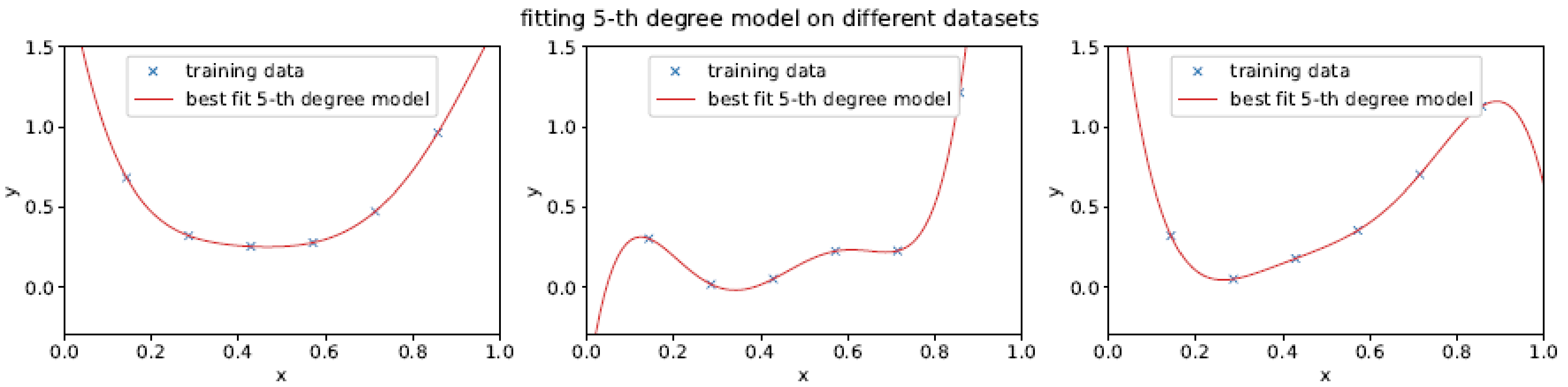
Bias-variance tradeoff
Example

- In this case, we fit patterns in the data that are present in our small, finite training set and do not reflect the wider pattern of the relationship between the input and outputs.
  - Spurious patterns resulted from noise



Bias-variance tradeoff
Example

- The **variance**: the amount of variations across models learnt on multiple different training datasets (drawn from the same underlying distribution).



Bias-variance tradeoff

- Often, there is a tradeoff between bias and variance.
  - Model is too simple and has very few parameters: it may have large bias (but small variance), and it typically may suffer from underfittng.
  - Model is too complex and has very many parameters: it may suffer from large variance (but have smaller bias), and thus overfitting

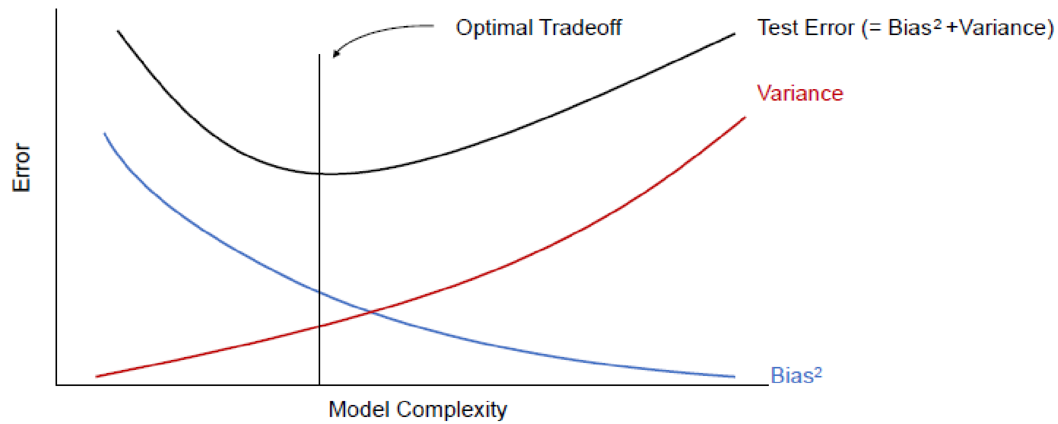


Figure 8.8: An illustration of the typical bias-variance tradeoff.