Machine Learning

| Main Challenges – Bad Data

Prof. Dr. Mostafa Elhosseini

Professor of Smart Systems Engineering

https://youtube.com/drmelhosseini

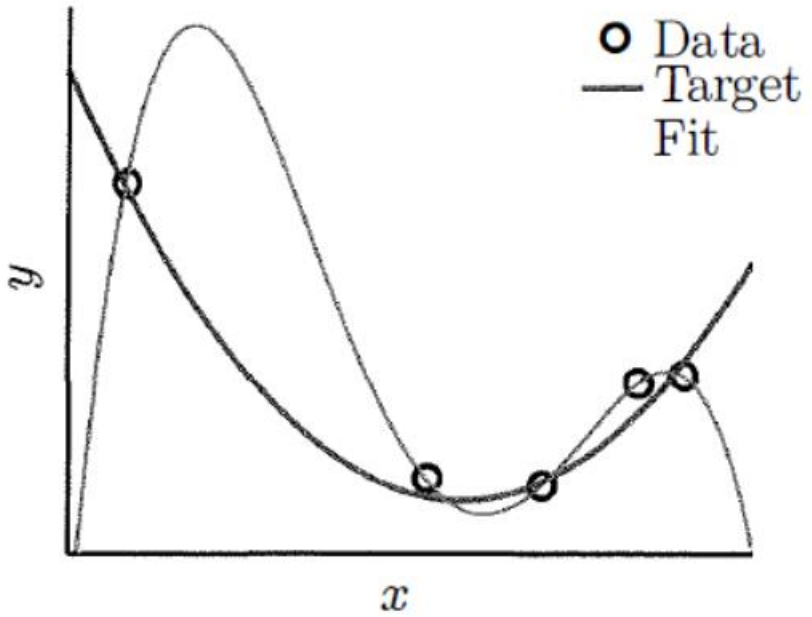
Main Challenges

- Since your main task is to select a model and train it on some data, the two things that can go wrong are:
  - bad model and
  - bad data

Bad Data

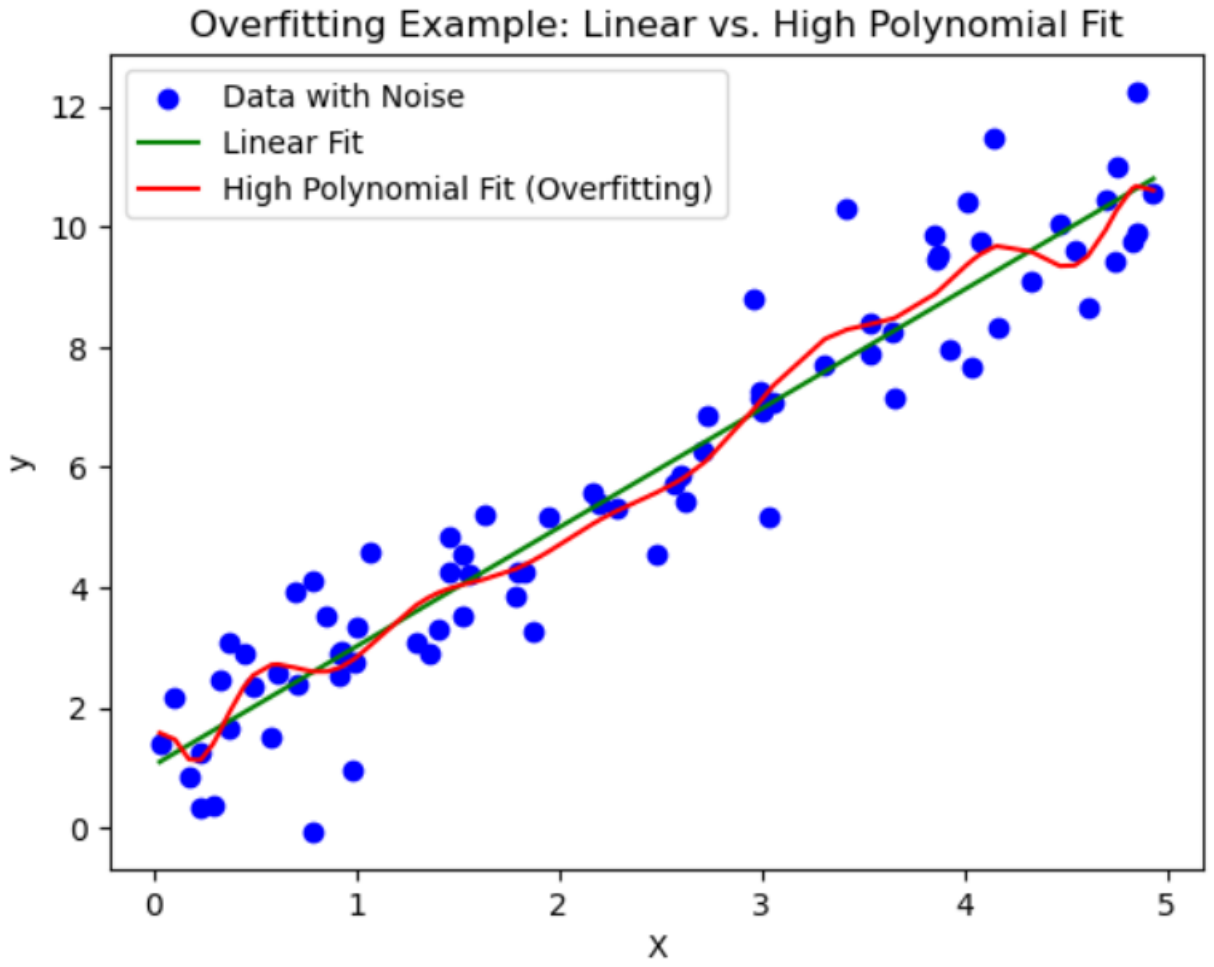
Overfitting:

- Overfitting occurs when a model learns the training data too well, capturing noise or random fluctuations in the data rather than the underlying patterns
  - Model performs well on the training data, but it does not generalize well
- Complex model uses its additional degrees of freedom to 'learn' the noise
- The little noise in the data has misled the learning



Bad Data

```
# Generate random data with a linear relationship
X = np.sort(5 * np.random.rand(80, 1), axis=0)
y = 2 * X.squeeze() + 1 + np.random.randn(80)
```



Bad Data

Key characteristics of overfitting include:

- High Training Accuracy: The model achieves a very high accuracy on the training data.
- Poor Generalization: Despite high accuracy on training data, the model fails to generalize well to new, unseen data.
- Capturing Noise: Overfit models capture the noise or random variations present in the training data, mistaking them for meaningful patterns.
- Complex Models: Overfitting often occurs with complex models that have a large number of parameters or features, allowing them to memorize the training data.

Bad Data

Causes of Overfitting:

- **Too Complex Model**: Models with too many parameters relative to the amount of training data are prone to overfitting.
- **Insufficient Data**: When the training dataset is small, the model may find it easier to memorize the data rather than generalize from it.
- **Noise in Data**: If the training data contains noise or irrelevant patterns, the model may learn these as if they are genuine features.

Bad Data

Addressing Overfitting:

- **Simplifying the Model**: Use simpler models with fewer parameters or features to reduce complexity.
- **Regularization**: Apply regularization techniques to penalize overly complex models, discouraging them from fitting the noise.
- **Cross-Validation**: Use techniques like cross-validation to assess model performance on different subsets of the data, helping to identify overfitting.
- **More Data**: Increase the size of the training dataset to provide the model with a more representative sample.
- **Feature Selection**: Choose relevant features and eliminate irrelevant ones to reduce the risk of overfitting.

Bad Data

Addressing Overfitting:

- **Simplifying the Model**: Use simpler models with fewer parameters or features to reduce complexity.
- **Regularization**: Apply regularization techniques to penalize overly complex models, discouraging them from fitting the noise.
- **Cross-Validation**: Use techniques like cross-validation to assess model performance on different subsets of the data, helping to identify overfitting.
- **More Data**: Increase the size of the training dataset to provide the model with a more representative sample.
- **Feature Selection**: Choose relevant features and eliminate irrelevant ones to reduce the risk of overfitting.

Bad Data

Addressing Overfitting:

- **Regularization** is a technique used to prevent overfitting and improve the generalization ability of a model.
- Overfitting occurs when a model fits the training data too closely, capturing noise and irrelevant patterns that may not generalize well to new, unseen data. Regularization introduces a **penalty term** to the model's objective function, discouraging overly complex or large parameter values.

Bad Data

Addressing Overfitting:

- Imagine you're teaching a robot to perform a task. If you teach it exactly based on the examples you've seen, it might memorize those examples (overfitting) and perform poorly on new tasks. Regularization is like telling the robot not to overthink and stick to general rules, preventing it from memorizing every detail. It helps the model generalize better to new situations by avoiding overly specialized learning.

Bad Data

Addressing Overfitting: