Machine Learning (ML)

Chapter 2:

Statistical Learning

Regression function and Classification Problems

Saeed Saeedvand, Ph.D.

Outline

In this Chapter:

- ✓ Introduction to Statistical Learning
- ✓ Regression function
- ✓ Curse of dimensionality
- ✓ Introduction to Classification Problems.

Aim of this chapter:

- ✓ Understanding the main reason why we need to know about statistical learning. Then the concepts of regression function as underlying concept in ML. Finally discussing classification problems.

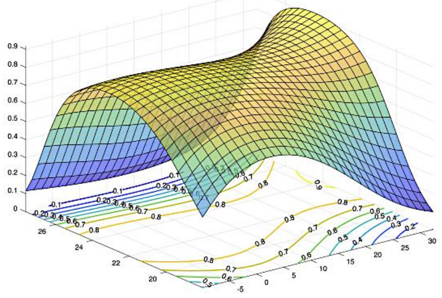
What is Statistical Learning?

- ✓ A branch of statistics and machine learning that focuses on developing and analyzing methods to make **predictions or decisions** based on data.

- ✓ The goal is to **build mathematical models** that can identify patterns in data and use these patterns to make predictions or decisions about new data.

- ✓ Statistical Learning is rooted in **mathematical theory** and **statistical inference** mostly.

- ✓ **Two main types of statistical learning:**
  - Supervised learning
  - Unsupervised learning



Comparison

Statistical learning and Machine learning

✓ Two **closely related fields** that **both deal with** the development of algorithms that can make predictions or decisions based on data.

Differences

- **Statistical learning** is a subfield of statistics that **focuses** on developing and analyzing methods for making predictions or decisions based on data.

- **Machine learning**, **on the other hand**, **includes** also **statistical learning** as well as **other approaches** to building algorithms that can learn from data.

- **So ML includes:**
  - ✓ Statistical models
  - ✓ Optimization algorithms
  - ✓ Deep learning
  - ✓ Neural Networks
  - ✓ …

Statistical learning vs Machine learning

Definition

- **Statistical learning** algorithms are often used in problems where the goal is to:
  - ➢ Understand the relationship between variables (e.g. regression analysis)

- **Machine learning** algorithms are often used in more complex problems:
  - ➢ Like image and **speech recognition**, **NLP**, and **anomaly detection**, etc.

- **Statistical learning** emphasis on interpretability.
- **Machine learning** emphasis on accuracy and is broader field.

Why Statistical Learning?

Why we need to know Statistical Learning?

- ✓ Although ML has powerful tools for building predictive models, but:
  - ❖ **Not a replacement** for understanding all the **underlying statistical concepts and principles**.

Why Statistical Learning?

Benefits:

- ✓ **Model selection and validation:**
  - ✓ Statistical learning helps you choose the best algorithm for a given problem and evaluate its performance.

- ✓ **Interpretability:**
  - ✓ Statistical learning provides more transparent and interpretable methods for analyzing data than some black-box machine learning models.

Why Statistical Learning?

Benefits:

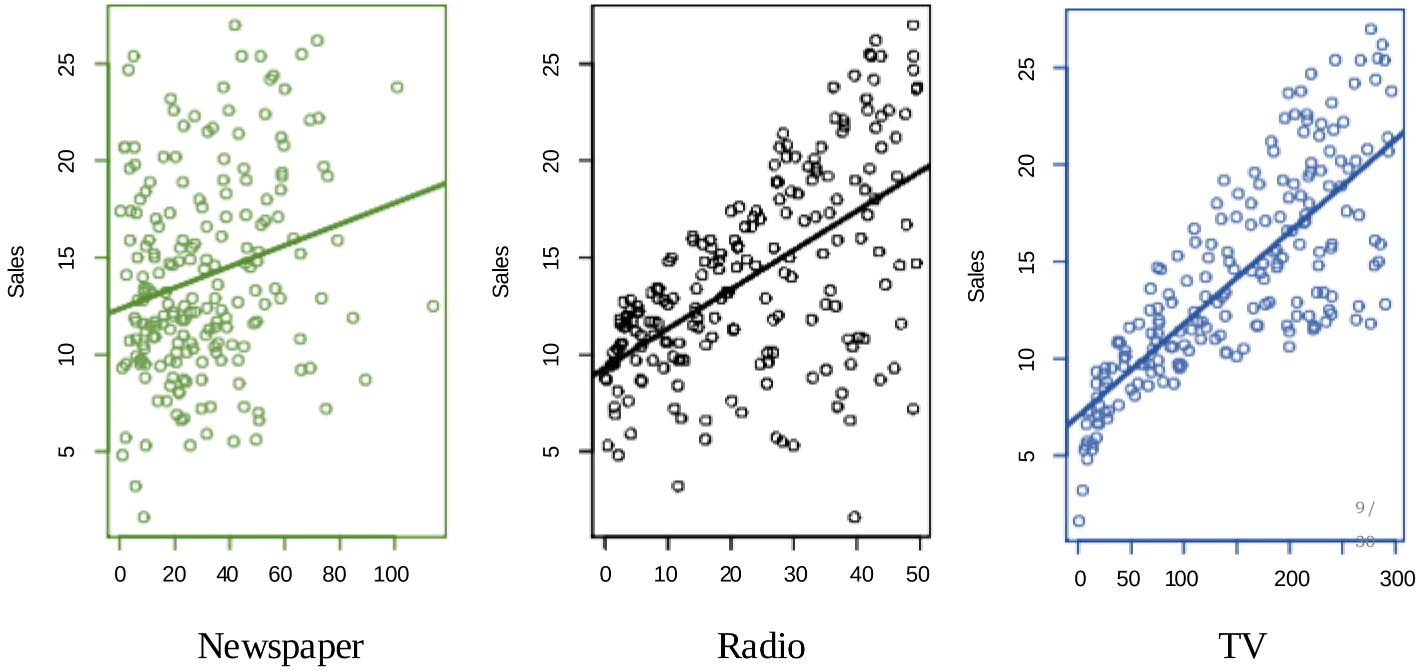
- ✓ **Data pre-processing:**
  - ✓ Statistical learning can provide techniques for **handling missing data**, **outliers**, and other issues that can affect the performance of ML algorithms.

- ✓ **Development of new algorithms:**
  - Understanding statistical learning is essential for developing and evaluating new algorithms.

Statistical Learning

Example

- ✓ Amount of Sales if we do advertisements on TV, Radio and Newspaper.



- ✓ The lines are linear-regression fit to each.

$$\text{Sales} \approx f(\text{Newspaper}, \text{Radio}, \text{TV})$$

Statistical Learning - Notations

- ✓ The goal is to predict Sales, (commonly we refer as Y).

- ✓ On the other hand, advertisements are an input variables labeled:
  - X1, X2, X3 (known as features or predictors).

- ✓ To refer to all the input variables together, we can use the term "**input vector**".

$$x = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

Statistical Learning - Notations

✓ Therefore the model can be written as follows:

$$y = f(X) + \varepsilon$$

Noise in the output variable that cannot be explained by the input variables X
e.g. price of the house is **not recorded accurately (y)**

Can we improve it?

Increase the sample size

Statistical Learning - Notations

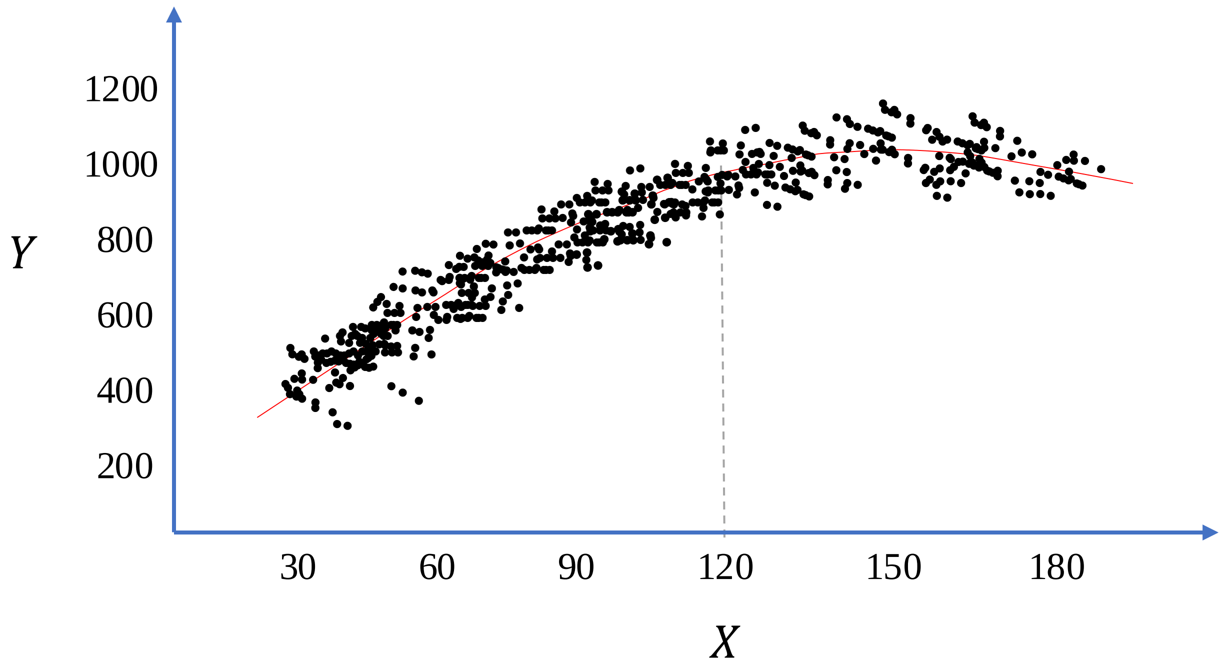
$$y = f(X) + \varepsilon$$

- ✓ If we have a **good *f*** we can make **better predictions of *Y*** at **new data points**.
- ✓ We need to **determine which features in** $X = (X_1, X_2, \dots, X_n)$ in explaining Y as output are important.
- ✓ For example Years of Education greatly influence Income, while Marital Status usually has little effect.

What is the Regression function?

Does an optimal f(X) exist?

- ✓ What is a suitable $f(x)$ value for a given X value, such as $X = 120$?
- ✓ There may be multiple Y values corresponding to it!



Regression function:

$$f(x) = E(Y|X = 120)$$

expected value (average) of Y given $X = 120$

Regression function

✓ We can define Regression function $f(x)$ for vector X:

$$f(x) = f(x_1, x_2) = E(Y|X_1 = x_1, X_2 = x_2)$$

We should minimize the **mean-squared prediction error** for all points $X = x$ for predicting Y over all functions f.

$$f(x) = E[(Y - f(X))^2|X = x]$$

Regression function

✓ We can calculate the error by:

$$\varepsilon = Y - f(x)$$

$$var(\varepsilon) = \varepsilon$$

Amount of variability in the dependent variable

✓ Error here called irreducible $(var(\varepsilon))$:

- Even if $f(x)$ is the best possible estimate noise cannot be predicted or explained by the model (e.g. for same input we have two different results already in dataset).

Regression function

✓ Consider $\hat{f}(x)$ as an estimation of the $f(x)$ we can write:

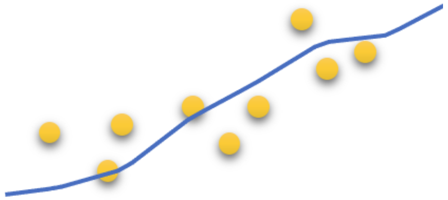
Reducible ***Irreducible***

$$f(x) = E\left[\left(Y - \hat{f}(x)\right)^2 | X = x\right] = [f(x) - \hat{f}(x)]^2 + var(\varepsilon)$$

True function

Reducible

- ✓ Can be reduced by improving the accuracy of the prediction



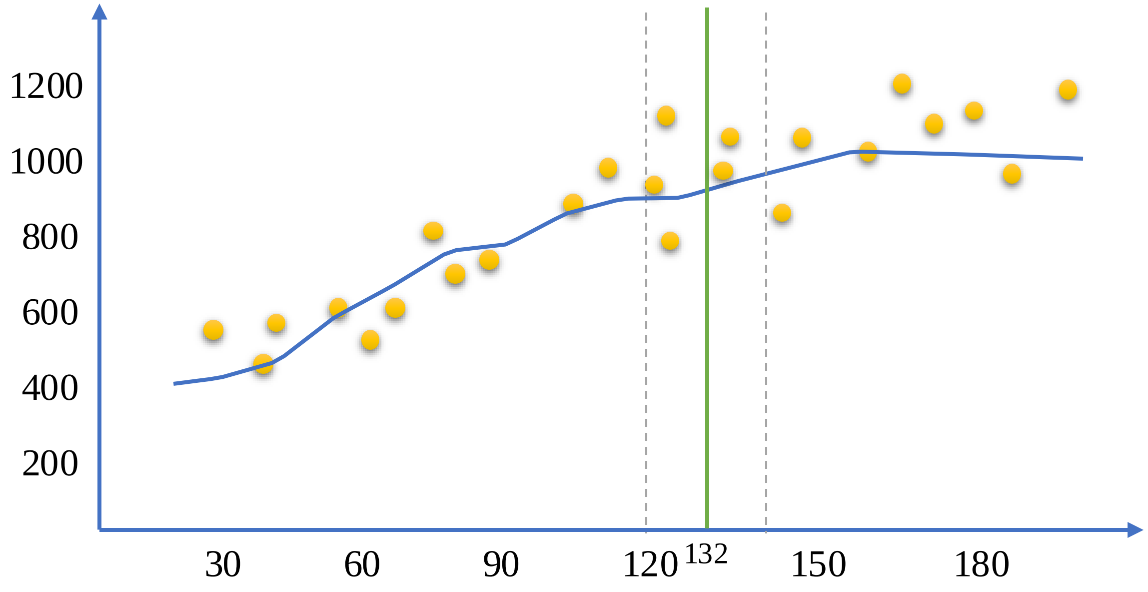
Question: Is this error reducible completely?

Regression function

How we can estimate f

✓ Most of times we don't have enough data points (like X =132) and computing $E[Y|X = x]$ **is not feasible**.

Solution?



We can relax the definition:

$$\hat{f}(x) = Avg(Y|X \in N(x))$$

Neighborhoods of x

Curse of dimensionality:

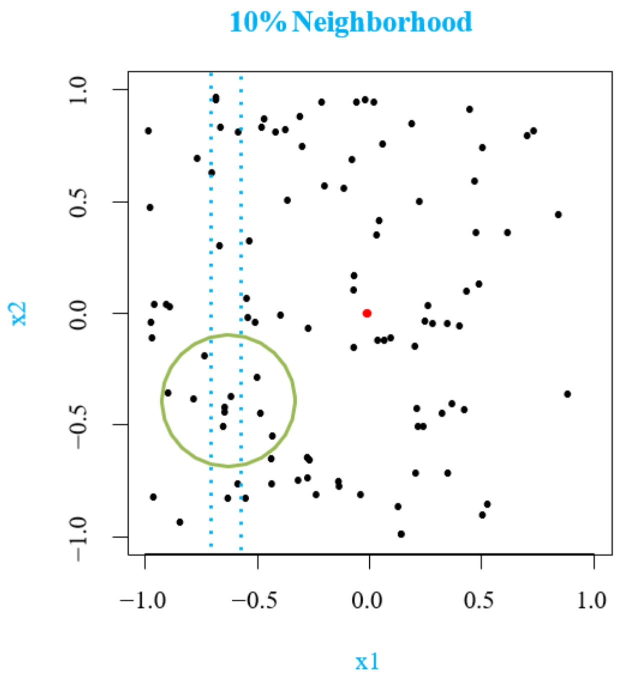
- ✓ If the dimensions is small (number of variables) **averaging neighborhoods** is good (e.g. less than 4), But in large dimensions it is a **poor approach** (called Curse of dimensionality).

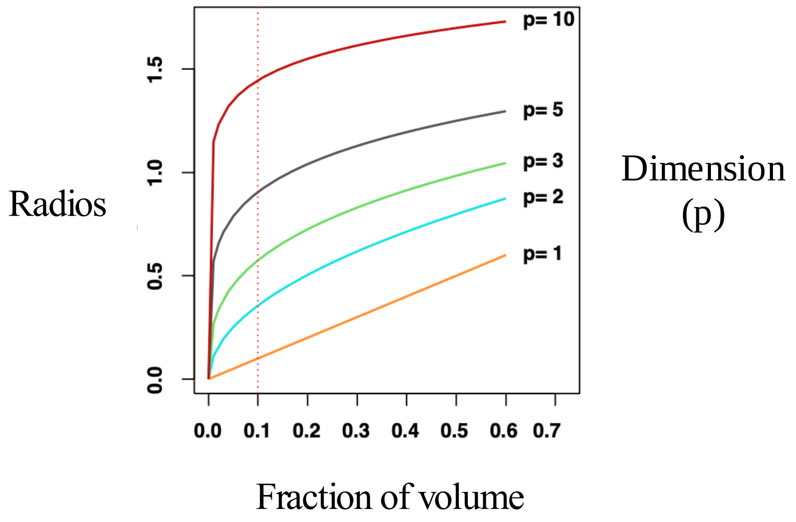
Curse of dimensionality problem:

- ✓ Curse of dimensionality is **dealing with high-dimensional data**.
- ✓ The **data can become increasingly sparse**, and the distance between any two data points becomes more and more similar.
- ✓ This makes it **challenging to analyze and model the data.**

Curse of dimensionality

- ✓ We need to have a **low variance** with having reasonable number of neighbors.
- ✓ In case that we have **large dimensions** to have lower variance we **have to engage more data** (like 10% of all) that is not good and not local anymore for predictions!





Parametric Models

- ✓ To avoid course of dimensionality challenge we use **parametric model.**
- ✓ For **parametric model** an important instance is the **linear model**.
- ✓ Linear model can be specified in terms of $p + 1$ parameters

$$f(x) = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p$$

- ✓ The process of fitting the model to training data happens by **estimating the parameters**.

Parametric Models

- ✓ The true function ***f(X)*** can be **approximated well** and can be **easily interpreted** by a simple **linear model**.

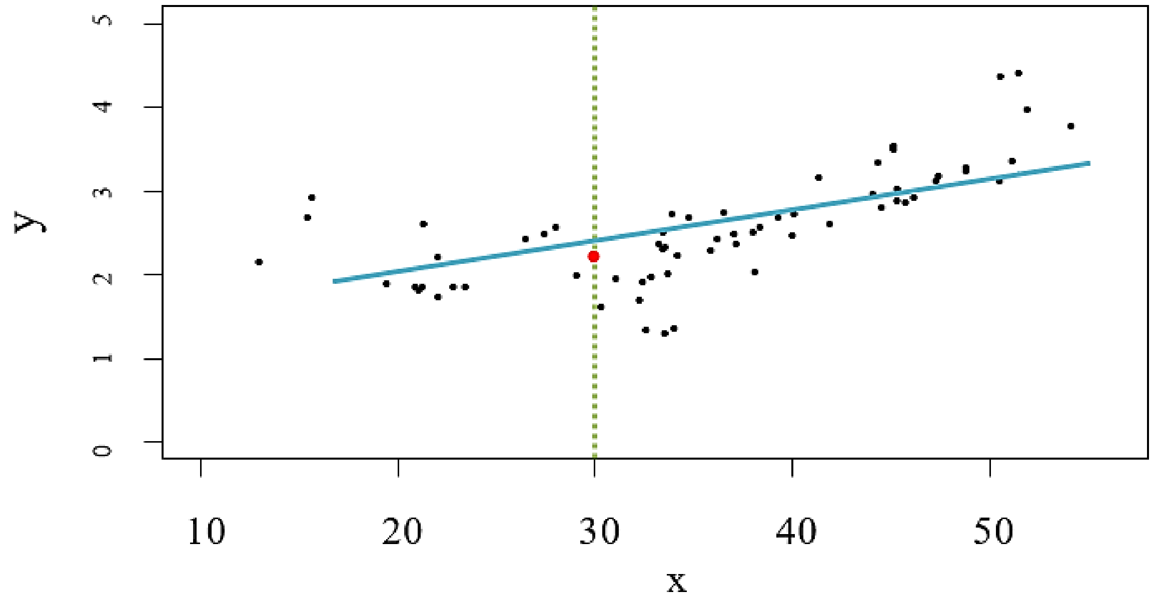
$$f(x) = \theta_0 + \theta_1 X_1 + \cdots + \theta_0 X_p$$

- ✓ linear model never correct and for complex data we will have large error.

Parametric Models

✓ For following example we can use a linear model:

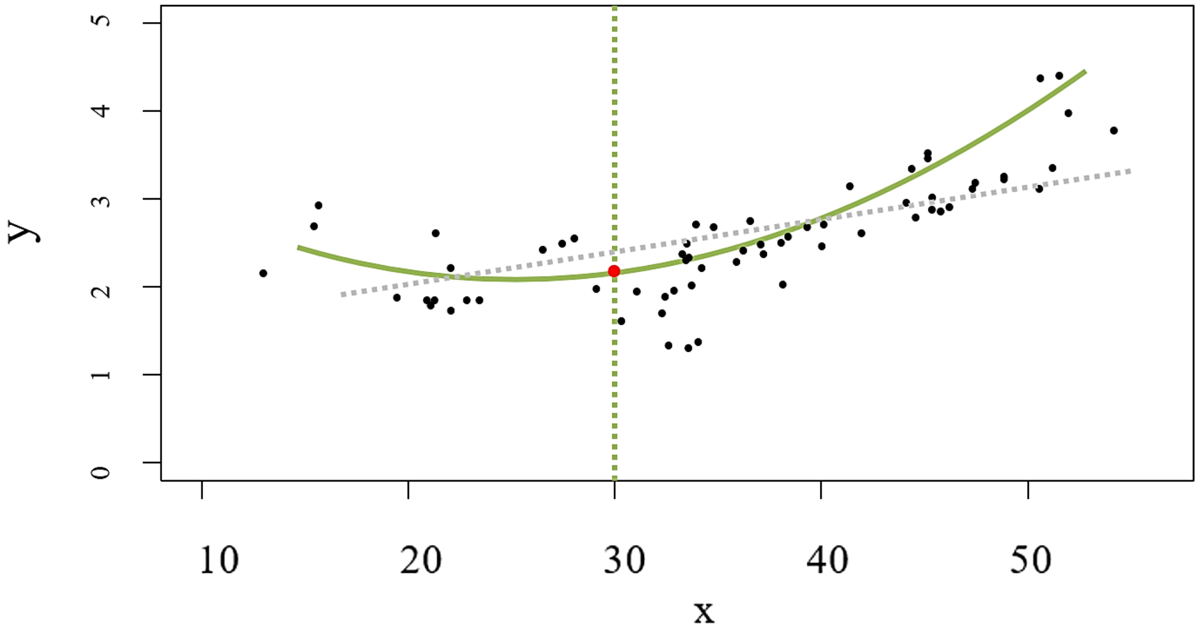
$$\hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 X$$



Parametric Models

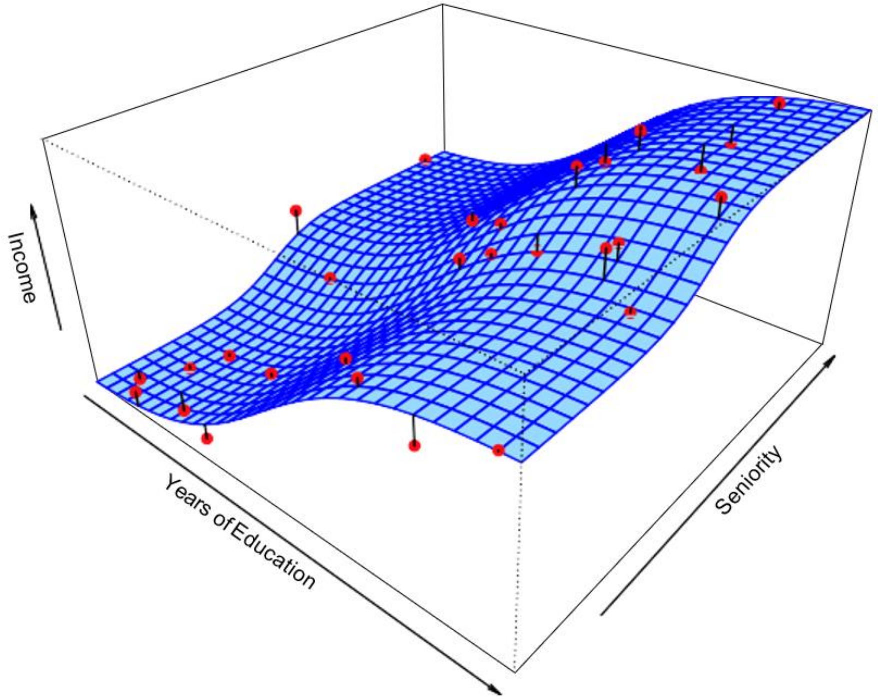
- ✓ Instead we can use a **quadratic model**, which can be better.

$$\hat{f}(x) = \hat{\theta}_0 + \hat{\theta}_1 X + \hat{\theta}_2 X^2$$



Example

A function f to estimate income based on education, and seniority.



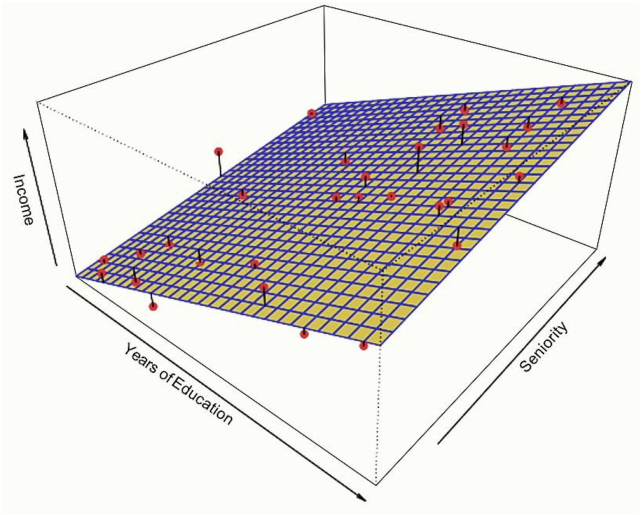
$$\text{income} = f(\text{education}, \text{seniority}) + \varepsilon$$

Example

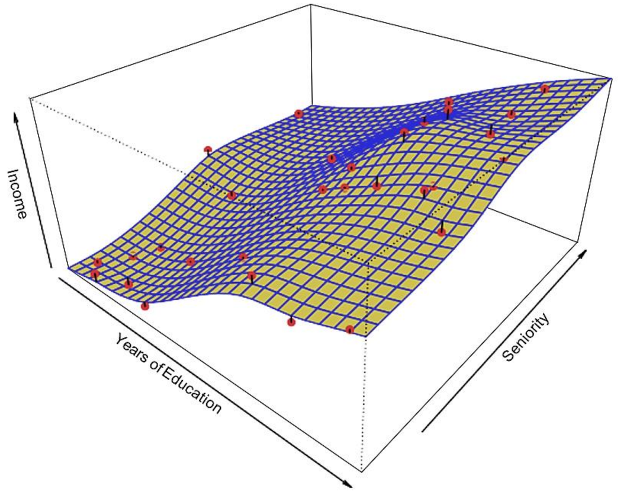
✓ We can write Linear regression model for this example:

$$\hat{f}(\text{education}, \text{seniority}) = \hat{\theta}_0 + \hat{\theta}_1 \times \text{education} + \hat{\theta}_2 \times \text{seniority}$$

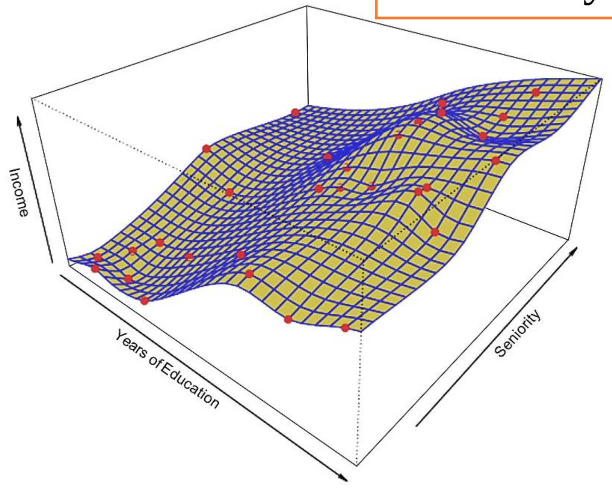
Is here any problem?



Linear regression model



Flexible regression model



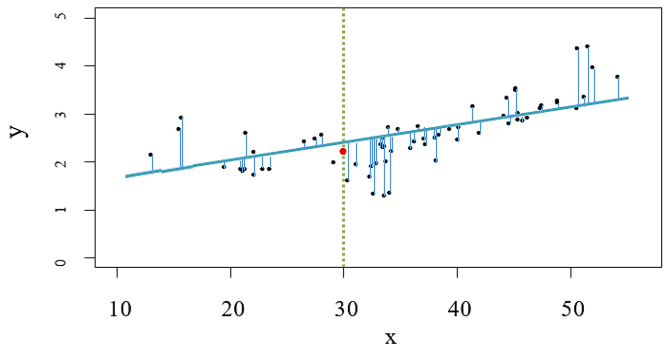
More Adjusted regression model

overfitting

The Model's Accuracy

- ✓ We can calculate the error of a model by computing the average squared prediction error for training data as Mean Square Error (MSE):

$$MSE_{Train} = \frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{f}(x_i)]^2, n = |train|$$

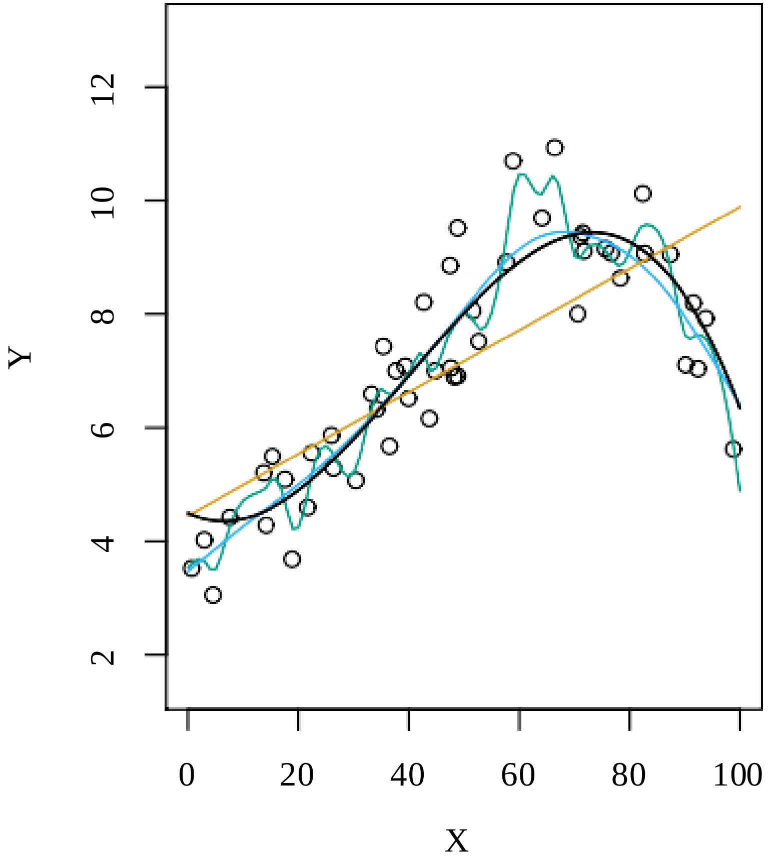


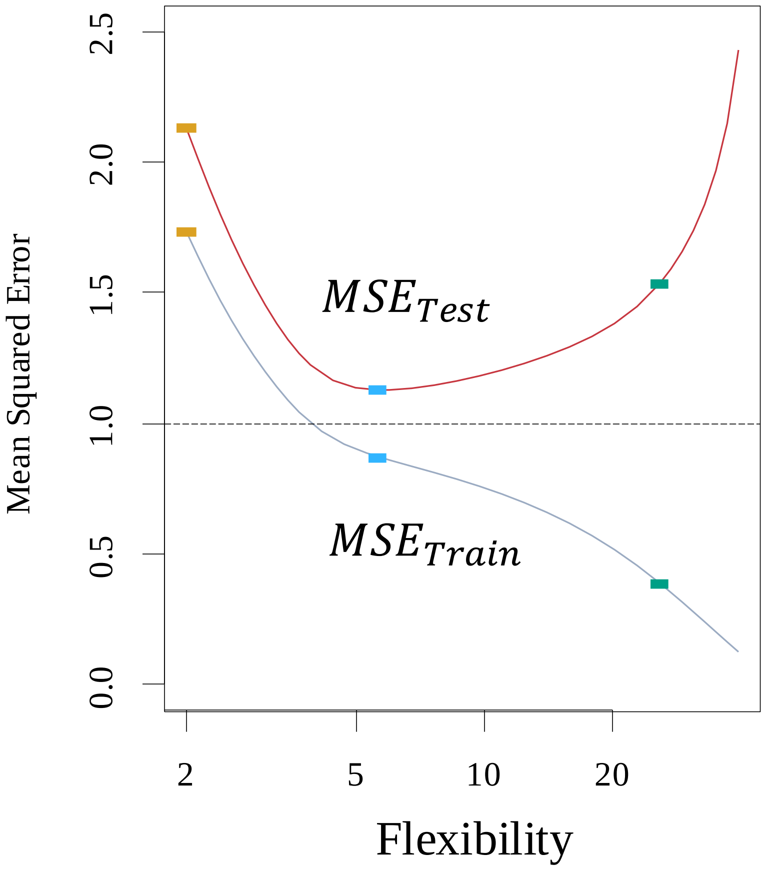
- ✓ And test data:

$$MSE_{Test} = \frac{1}{m}\sum_{i=1}^{m}[y_i - \hat{f}(x_i)]^2, m = |test|$$

Different flexibility of training model

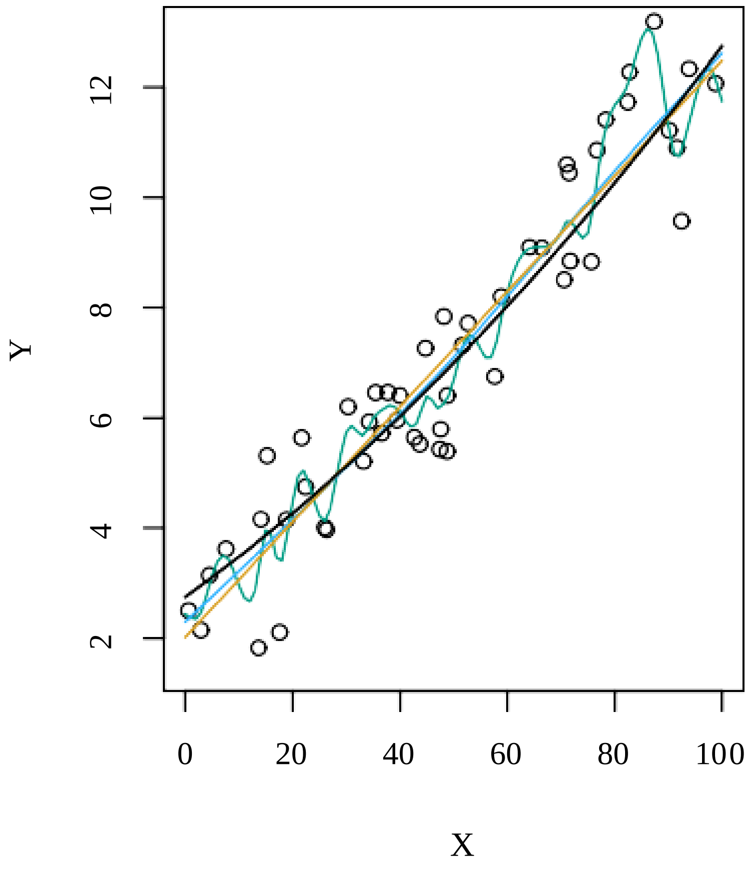
Fits of different flexibility

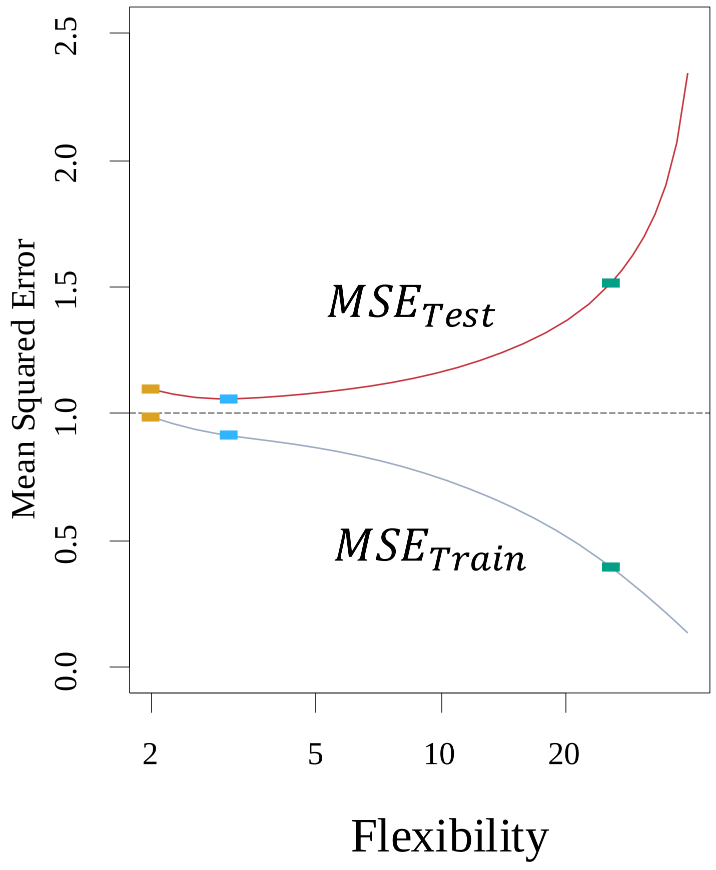




Different flexibility of training model

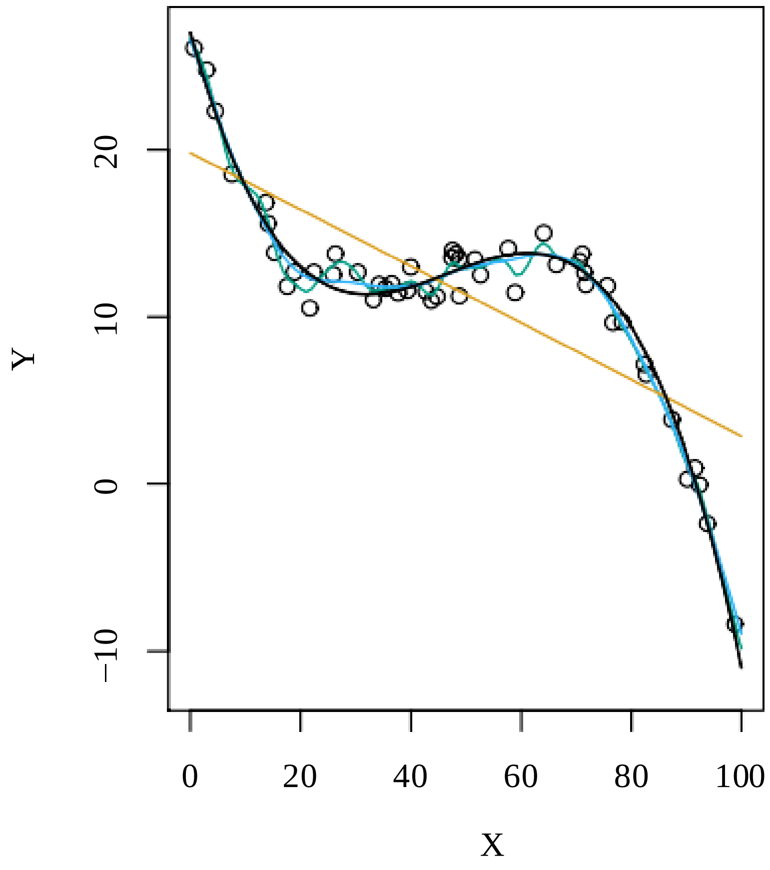
Fits of different flexibility

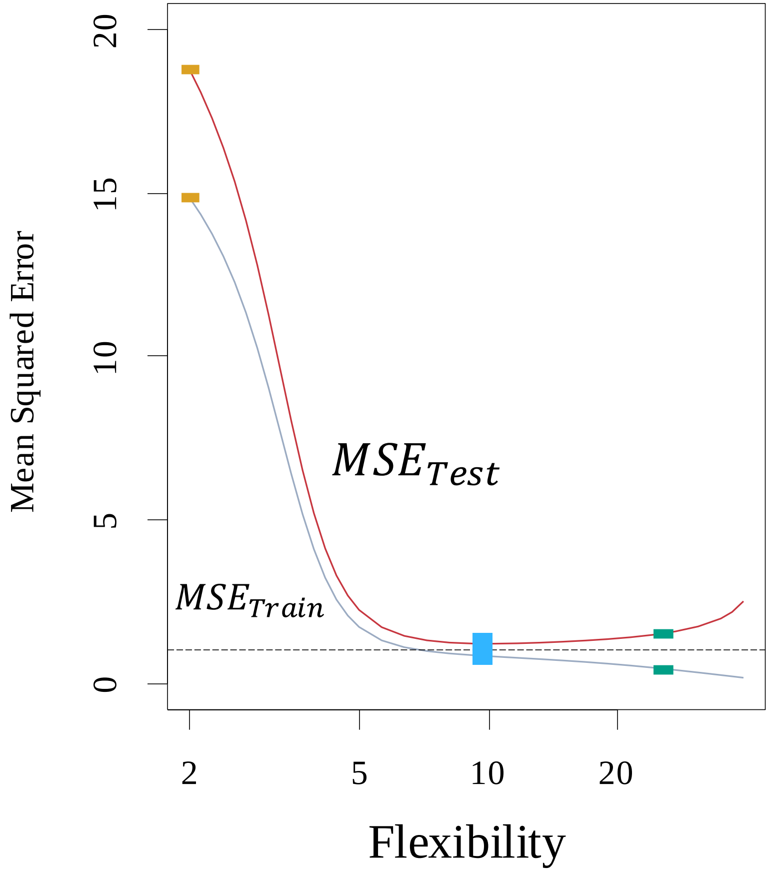




Different flexibility of training model

Fits of different flexibility





Different flexibility of training model

Flexibility of training model

✓ In ML and Statistical learning we aim to develop models that can predict the output for new, unseen inputs with accuracy.

✓ A model's prediction error can be divided into two components:

$$E\left[(Y - \hat{f}(x))^2\right] = bias^2 + Var + \varepsilon$$

Bias:

- **High bias:** Shows the error from oversimplifying a real-world problem, leading to underfitting.
- **Low Bias:** Shows the model is close to the true function.

$$bias(\hat{f}(\mathrm{x})) = E\left[\hat{f}(x)\right] - f(x)$$

true function (ground truth)

Expected value of the model's predictions

Different flexibility of training model

Flexibility of training model

Variance:

- **How much the predictions fluctuate** around the average prediction
  - **High Variance:** The model is too sensitive to training data (predictions fluctuate a lot and overfitting).
  - **Low Variance**: The model is stable and produces (similar predictions for different training sets).

- Variance is the error from overcomplicating a model, leading to overfitting.

$$Var(\hat{f}(x)) = E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]$$

Predicted value of the target variable for a given input x

Expected value of the predicted values over **all possible values of x** (mean).

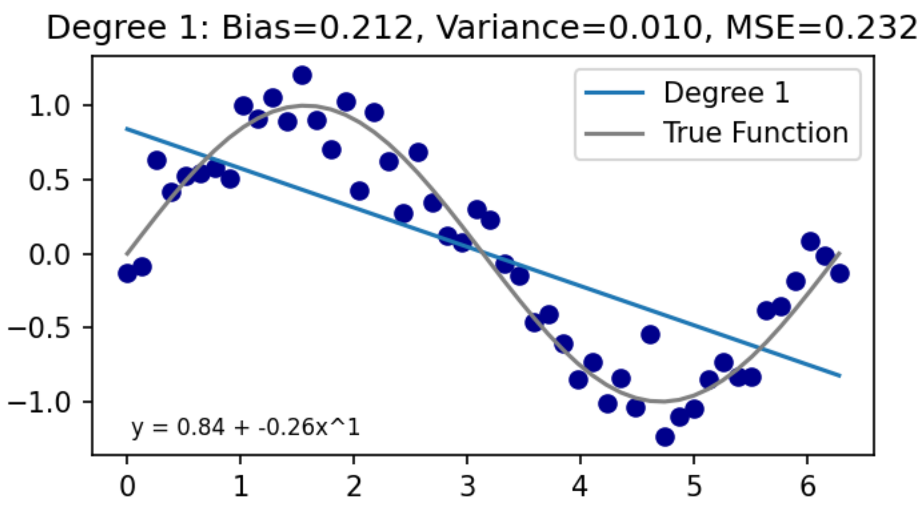
Different flexibility of training model

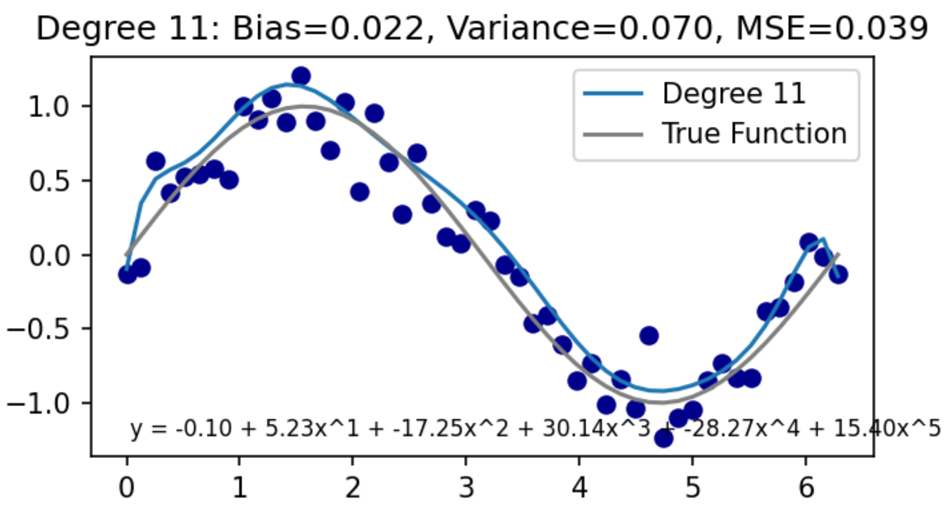
Bias-variance tradeoff

- ✓ **Bias-variance tradeoff** refers to the balance between the simplicity and flexibility of a model.
- ✓ Models that are **too simple have** low variance but high bias, while models that are **too complex have** high variance but low bias.
- ✓ The **goal** is to find the sweet spot where the bias and variance are balanced, resulting in a model that generalizes well to new data.

Different flexibility of training model

Bias-variance tradeoff





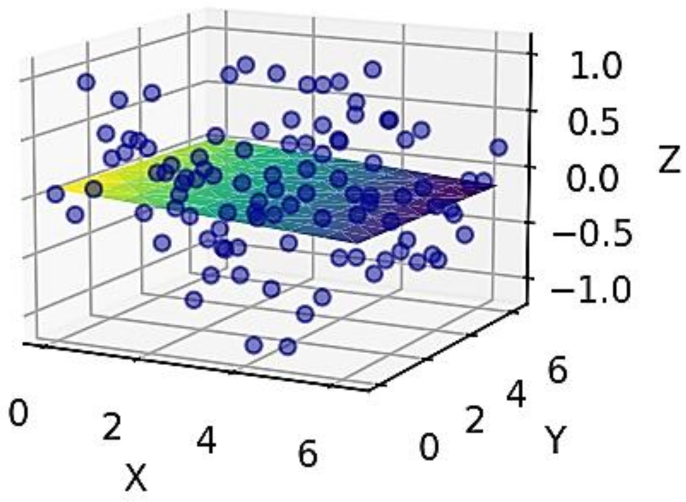
Python Example

Assignment

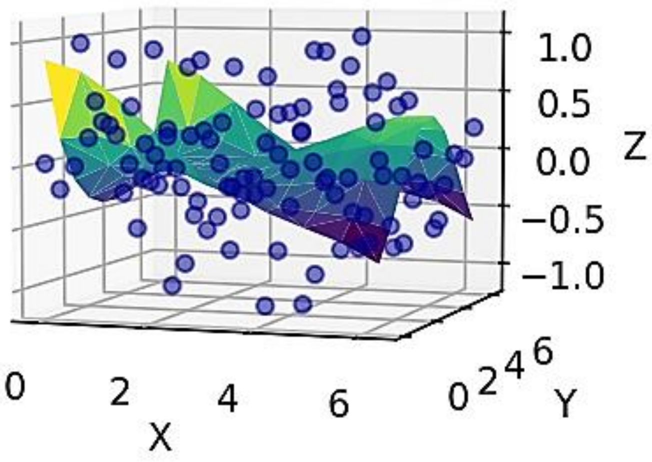
- ✓ A) Create a simple dataset with two variables as price and size for house (100 samples). You can write a loop with adding semi-random value to create your own library (data needs to have meaningful relationship. Then Plot the data with labels on each axes using matplotlib in python.
- ✓ B) Extend the same code to add one more variable as “location grade” and plot separately (2D).

Different flexibility of training model

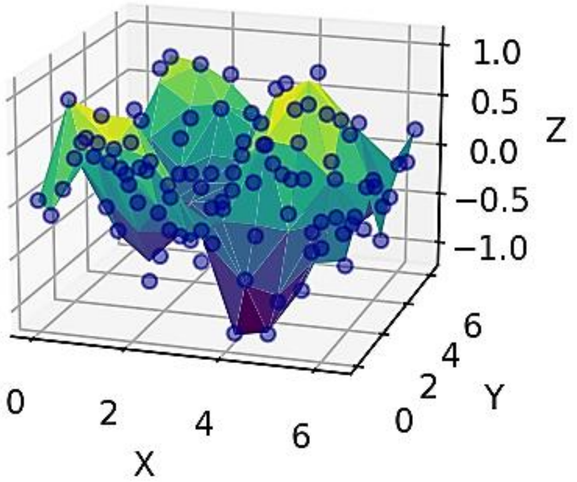
Bias-variance tradeoff



Degree 1: Bias=0.247, Variance=0.003, MSE=0.240



Degree 4: Bias=0.166, Variance=0.072, MSE=0.167



Degree 11: Bias=0.032, Variance=0.226, MSE=0.013

High bias

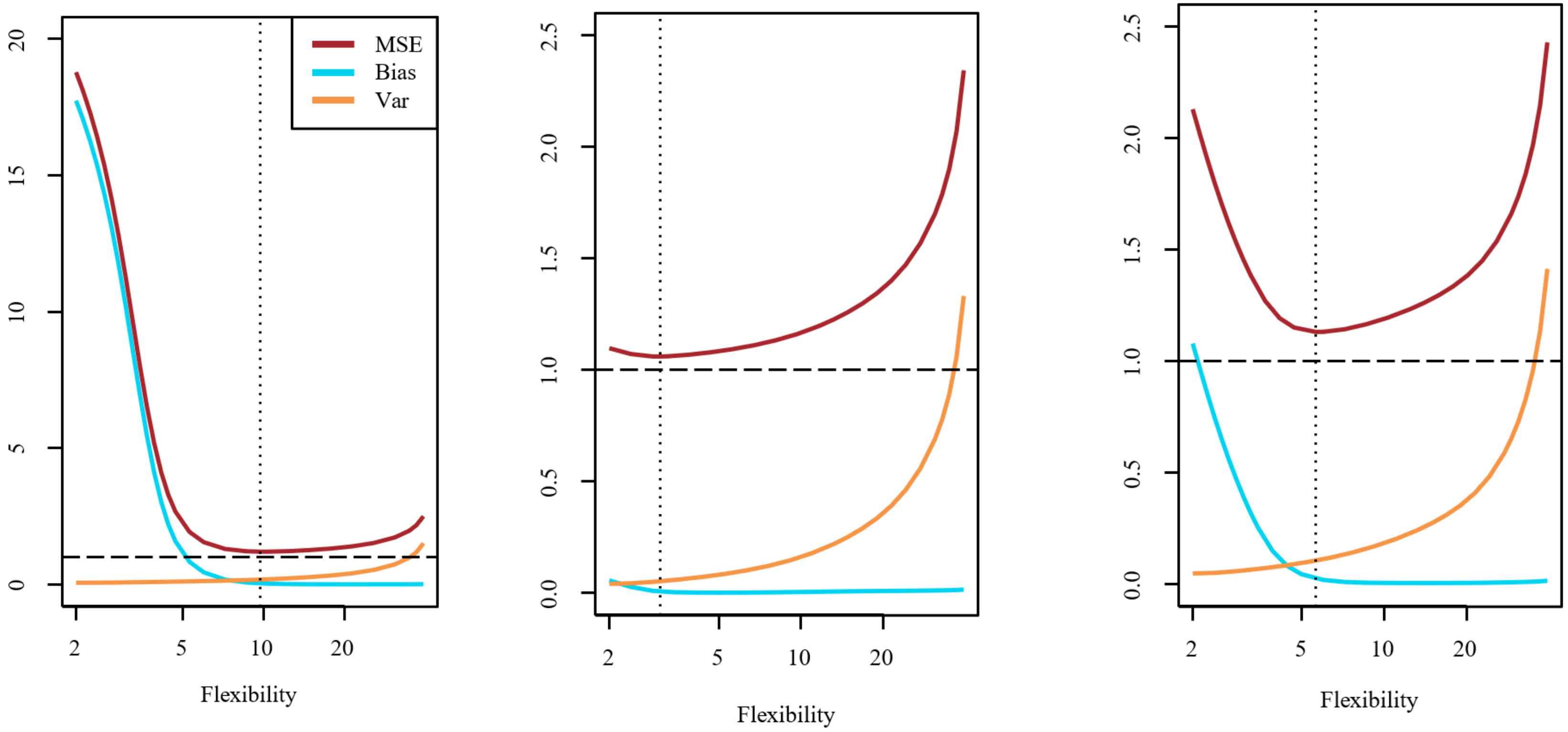
Low variance

Low bias

High variance

Different flexibility of training model

Bias-variance tradeoff



Classification

Classification Problems

- ✓ Type of ML problem in which the goal is to predict the class or category of a **given input** based on a set of **labeled training data**.
- ✓ Learn a **mapping function** from input features to class labels.
- ✓ In **image classification**, an image is inputted, and the classes could be different types of objects or animals, such as cats, dogs, or birds.

Classification

What is the Conditional class probabilities?

✓ Probability of a specific class given an input value x.

$$P_k(x) = Pr(Y = k|X = x), where\ k = 1, 2, .., K$$

Probability of belonging input x to class k

class of variable

✓ These probabilities can be used to **classify new, unlabeled observations.**

✓ We need to trained ML models, which has learned to **map input features** to **class of labels** based on labeled training data.

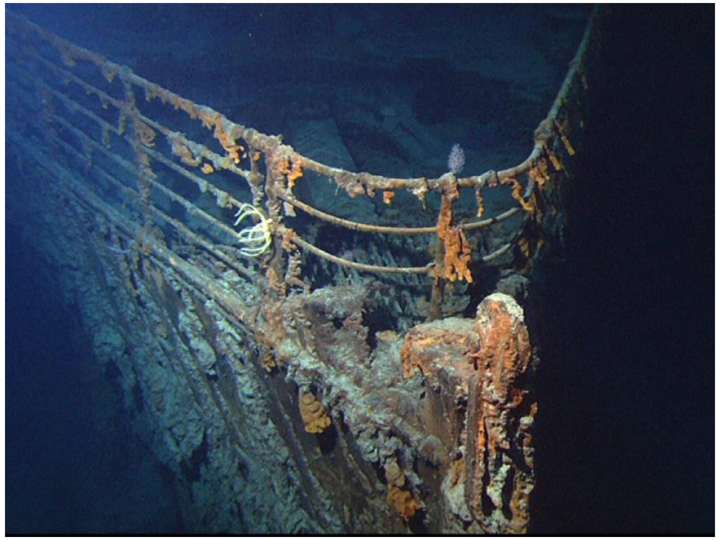
Classification

Bayes' theorem

- ✓ Bayes' theorem is widely used in ML and data science.
- ✓ The most important concept in probability theory to model and reason uncertainty.
- ✓ In **1998,** Tommy Thomson, et.al used it to uncover a ship that sunk in century (worth 50,000,000$).
- ✓ By incorporating multiple sources of information into a probabilistic model, and prioritize their search efforts.

Assignment

What parameters they considered? formulate



Classification

Bayes' theorem

- ✓Describes the relationship between conditional probabilities.
- ✓Probability of a output y given some observed evidence x:

$$P(y|x)$$

Bayes tells us how to update our belief for new inputs

Summery

- ✓ We discussed Statistical Learning vs Machine learning
- ✓ We saw what is the Regression function
- ✓ We understood the curse of dimensionality concept
- ✓ We explained classification Problems idea