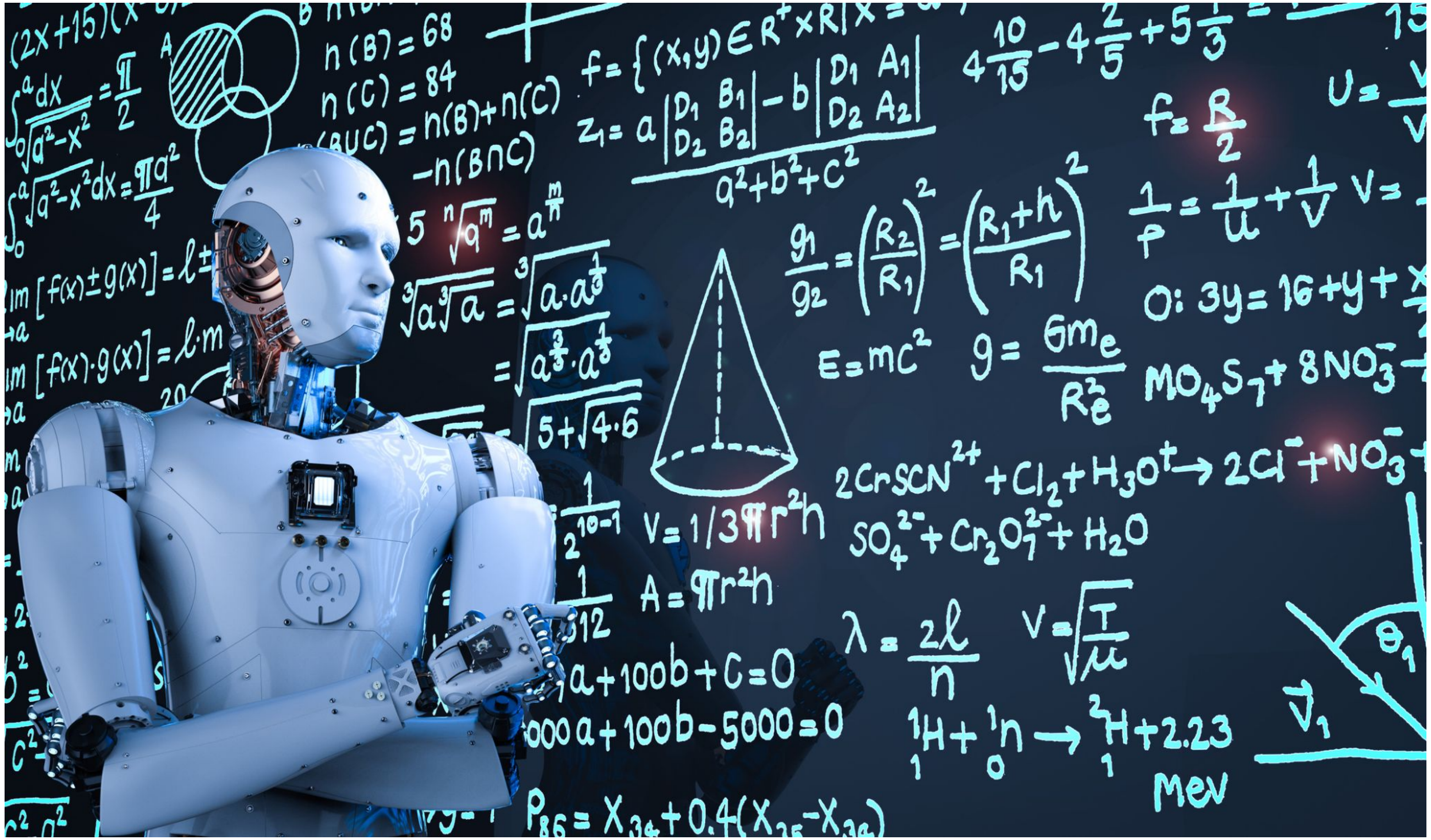
Diagnostics for Machine Learning Model



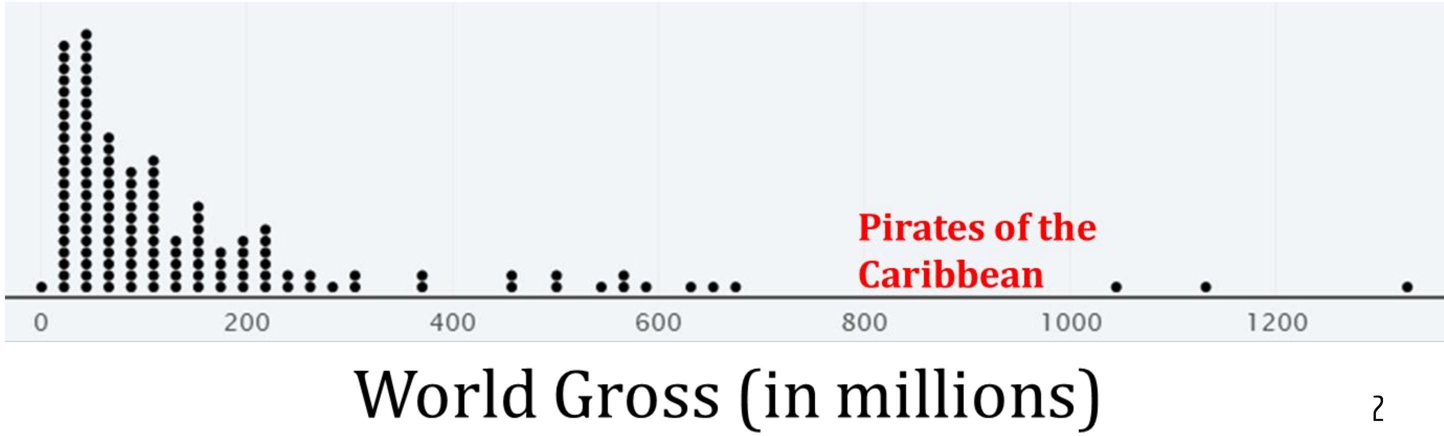
PhD. Msc. David C. Baldears S.
PhD(s). Msc. Diego Lopez Bernal

Outliers

An outlier is an observed value that is notably distinct from the other values in a dataset.

Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

- The outlier is though as a mistake(normally not)
- Decide whether the outlier is part of your population of interest or not
- See how much the outlier(s) are affecting the results



Outlier

- Outliers are useful to detect significant deviations from normal behavior Applications:
    - Credit Card Fraud Detection
    - Network Intrusion Detection

Missing Values

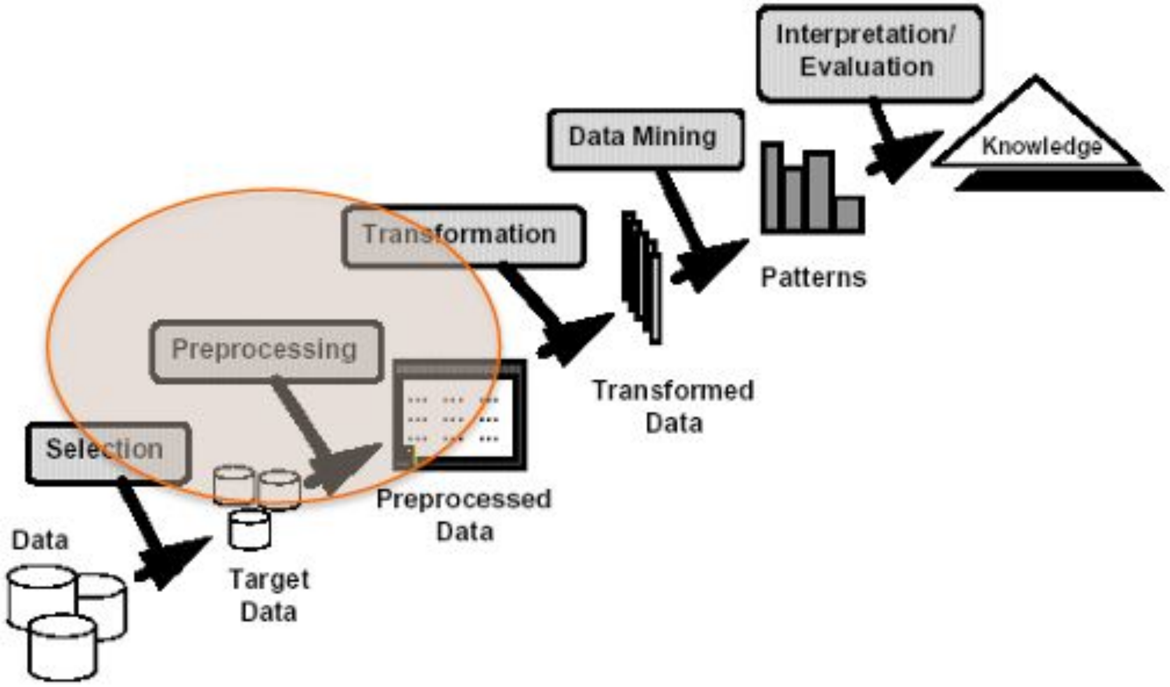
- Reasons for missing values
  - Information is not collected (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

Duplicate Data

- Data set may include data objects that are duplicates, or almost duplicates of one another
  - Major issue when merging data from heterogeneous sources
- Examples:
  - Same person with multiple email addresses
- Data cleaning
  - Process of dealing with duplicate data issues

Data Preprocessing

- Aggregation
- Sampling
- Dimensionality Reduction
- Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation



Aggregation & Sampling

Aggregation

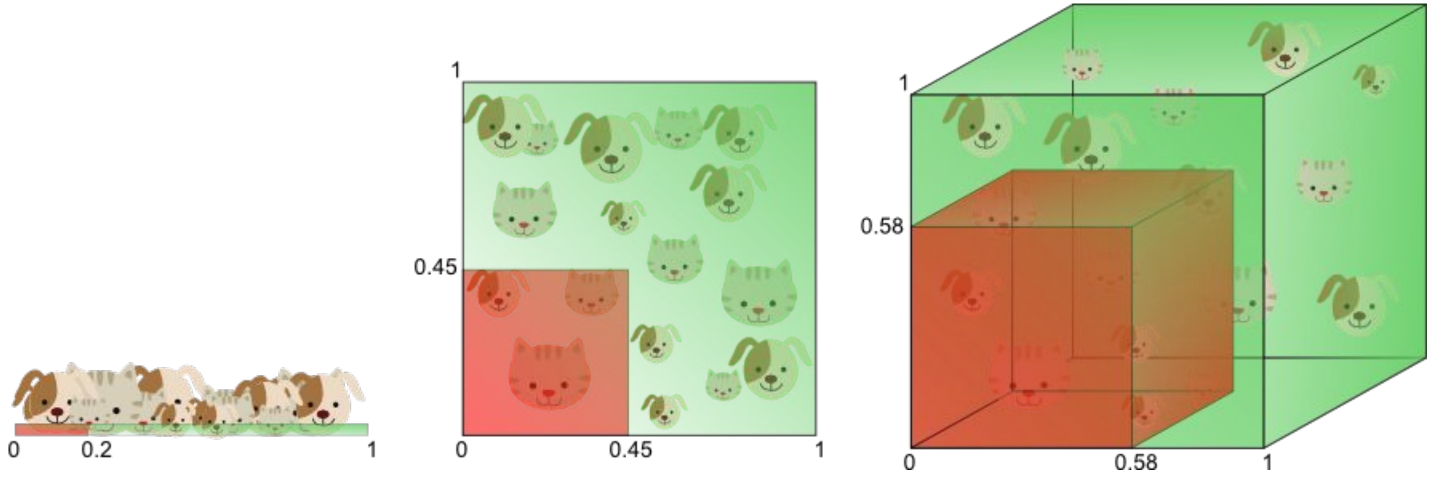
- Combining two or more attributes (or objects) into a single attribute (or object)
- Purpose
  - Data reduction Reduce the number of attributes or objects
  - Change of scale Cities aggregated into regions, states, countries, etc
  - More “stable” data Aggregated data tends to have less variability

Sampling

- Sampling is the main technique employed for data selection. – It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because processing the entire set of data of interest is too expensive or time consuming.

Curse Of Dimensionality

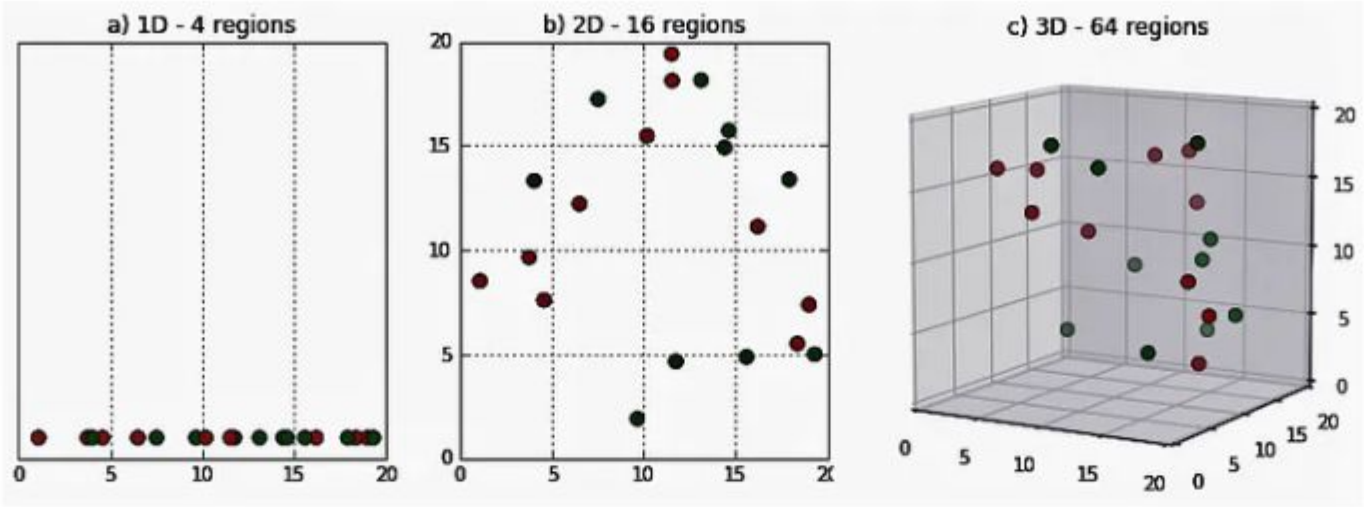
- Dimensionality: amount of features that describe our dataset
- When dimensionality increases, data becomes increasingly sparse in the space that it occupies



- The amount of training data to cover 20% of the feature space grows exponentially. In other words: more features need more data.

Curse Of Dimensionality

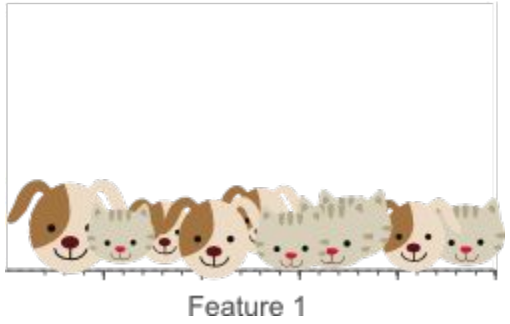
- Another way to observe this phenomenon:

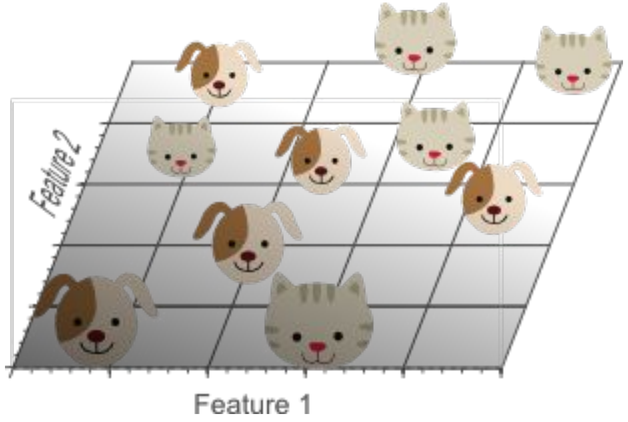


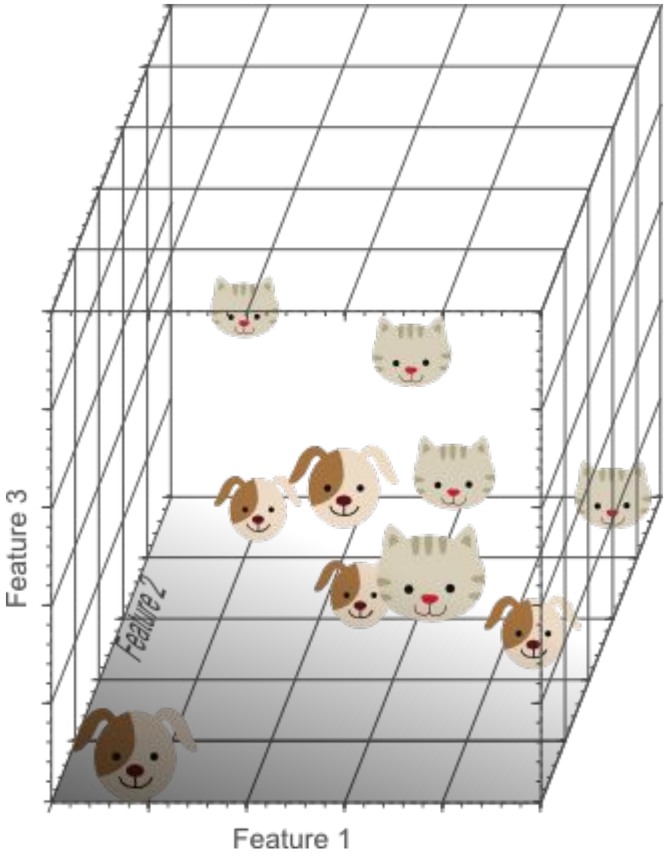
- Ass you add new dimensions, you create new space that is not filled with data. Therefore, you need more data for it to work well.
- Definitions of density and distance between points, which is critical for clustering and outlier detection, become less meaningful.

Curse Of Dimensionality

- It is also one of the main causes of overfitting:



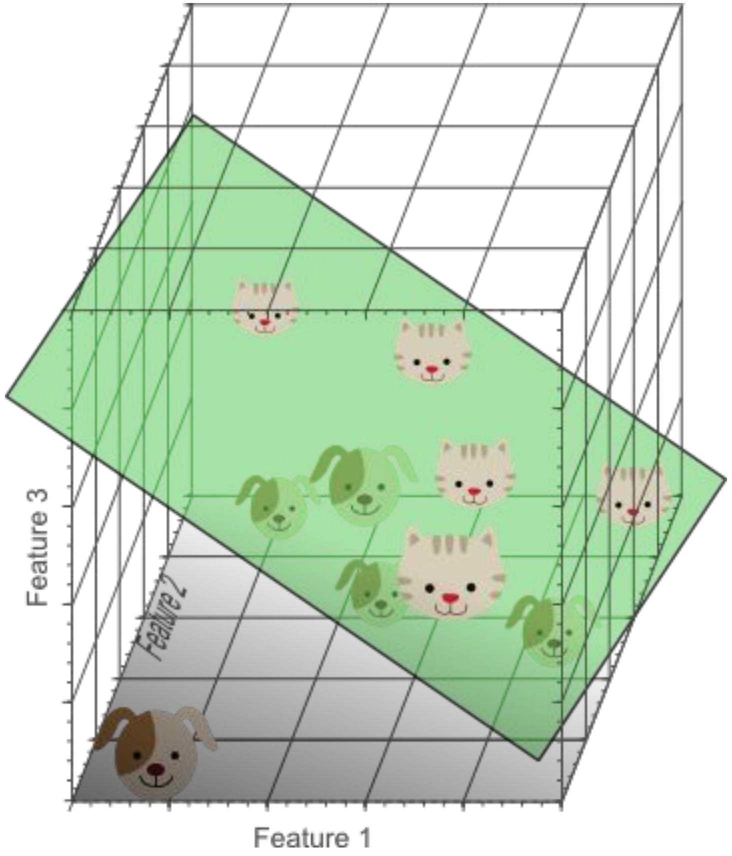




- Can we separate them?

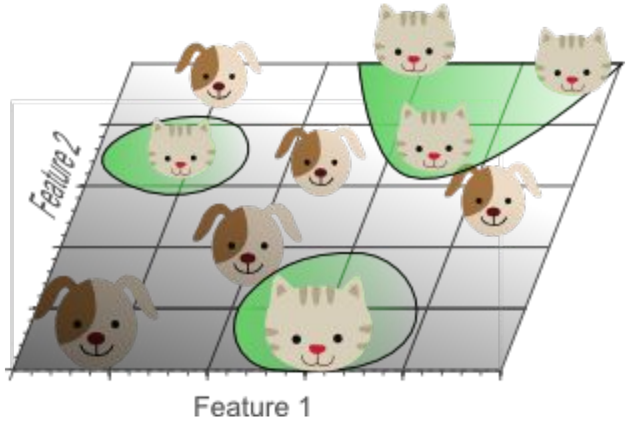
Curse Of Dimensionality

It is also one of the main causes of overfitting:

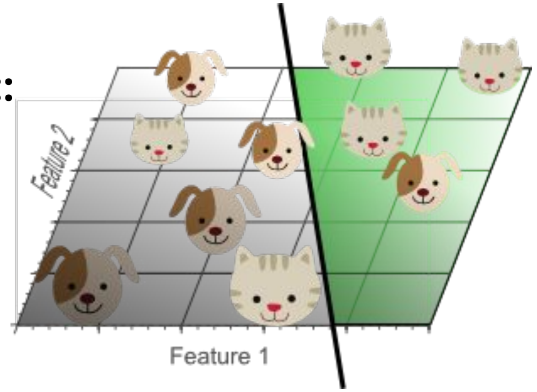


When projected into 2D:

Overfitting!



Better option::



Curse Of Dimensionality

- How to solve it?
- Dimensionality Reduction
  - Avoid curse of dimensionality
  - Reduce amount of time and memory required by data mining algorithms
  - Allow data to be more easily visualized
  - May help to eliminate irrelevant features or reduce noise
  - Avoid overfitting

Feature Subset Selection

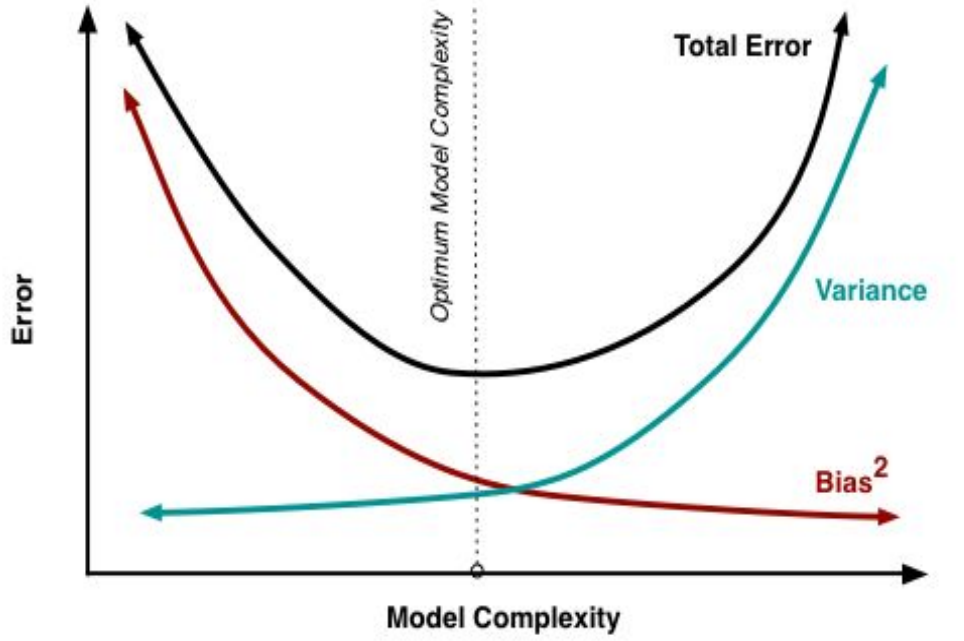
Reduce dimensionality of data

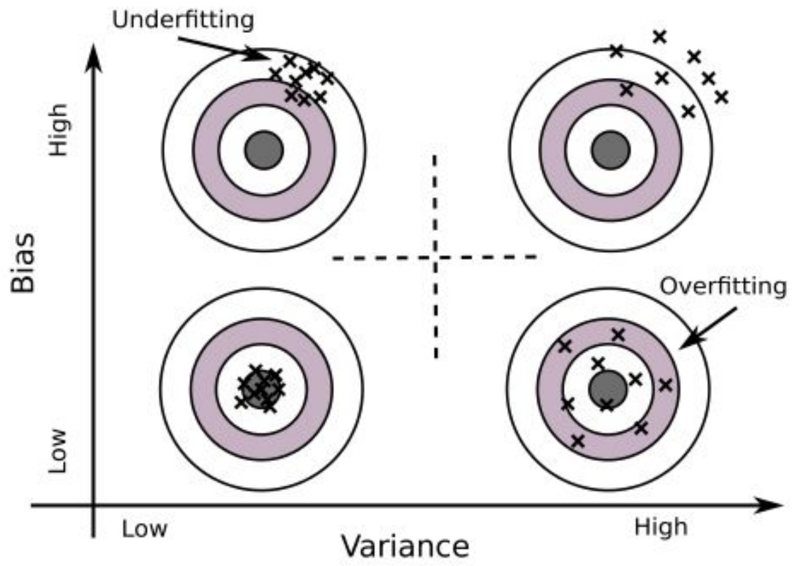
Remove:

- Redundant features
  - duplicate much or all of the information contained in one or more other attributes
  - Example: purchase price of a product and the amount of sales tax paid
- Irrelevant features
  - contain no information that is useful for the data mining task at hand
  - Example: students' ID is often irrelevant to the task of predicting students' GPA
- Techniques
  - Brute-force approach: Try all possible feature subsets as input to data mining algorithm
  - Embedded approaches: Feature selection occurs naturally as part of the data mining algorithm
  - Filter approaches: Features are selected before data mining algorithm is run

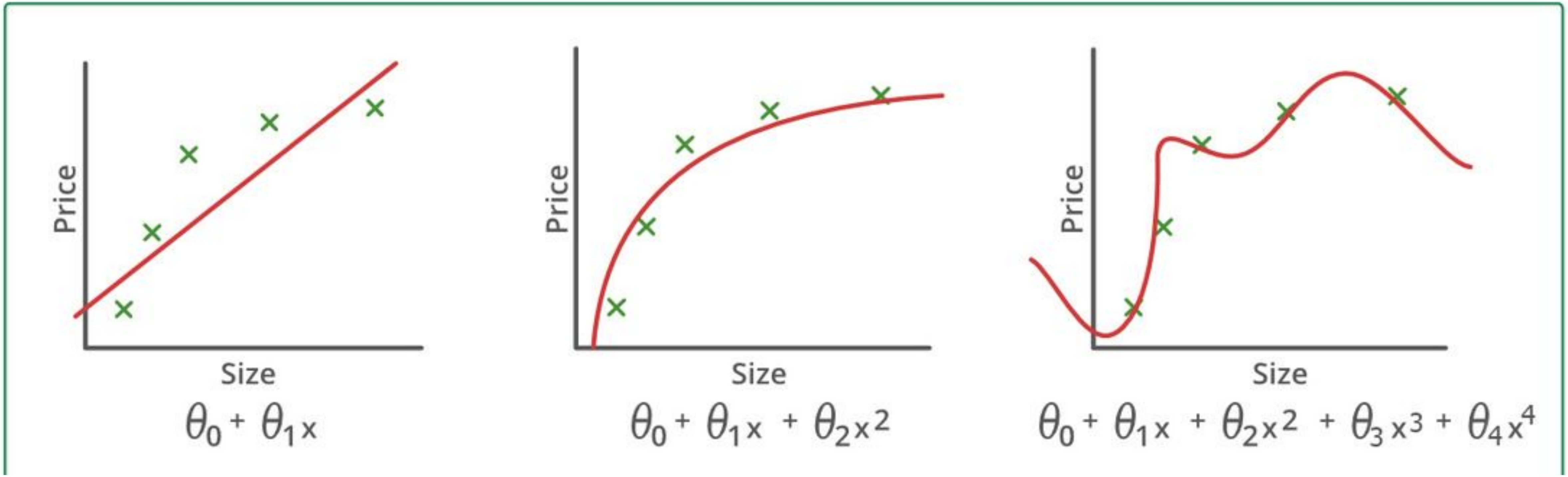
Bias and Variance

- Bias:
  - Assumptions made by a model to make a function easier to learn.
  - It is actually the error rate of the training data. When the error rate has a high value, it has High Bias
  - when the error rate has a low value, it has low Bias.
- Variance:
  - The error rate of the testing data is called variance.
  - When the error rate has a high value, it has High variance
  - When the error rate has a low value, it has Low variance.





Bias and Variance



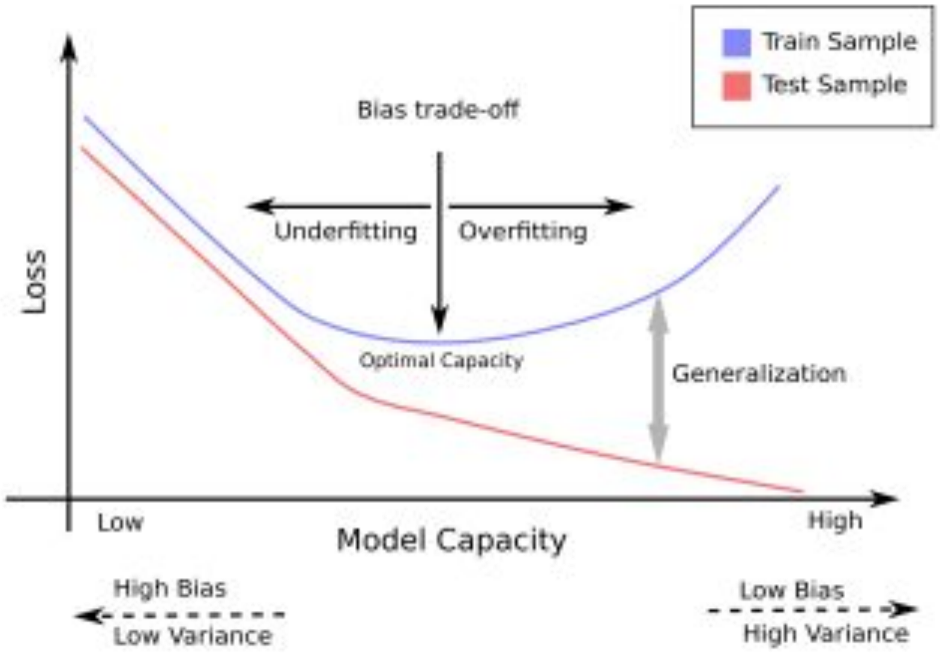
High bias (underfit)

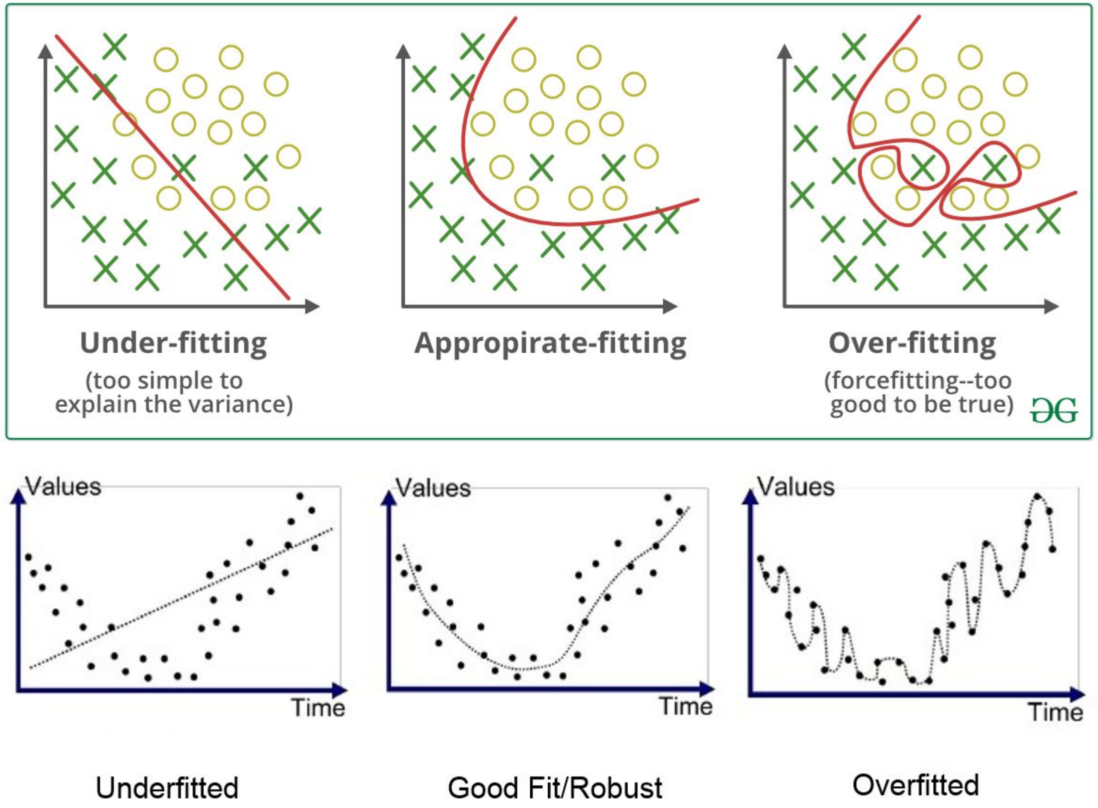
High bias (underfit)

High variance (overfit)

Model Complexity

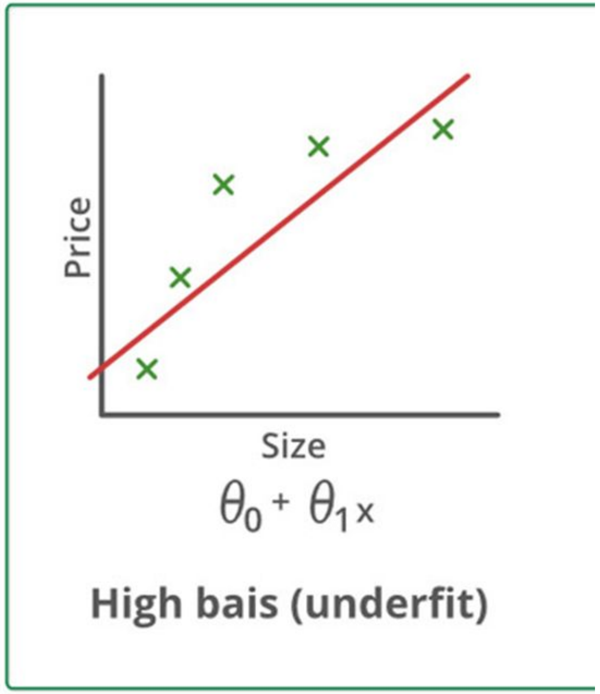
Underfitting and Overfitting



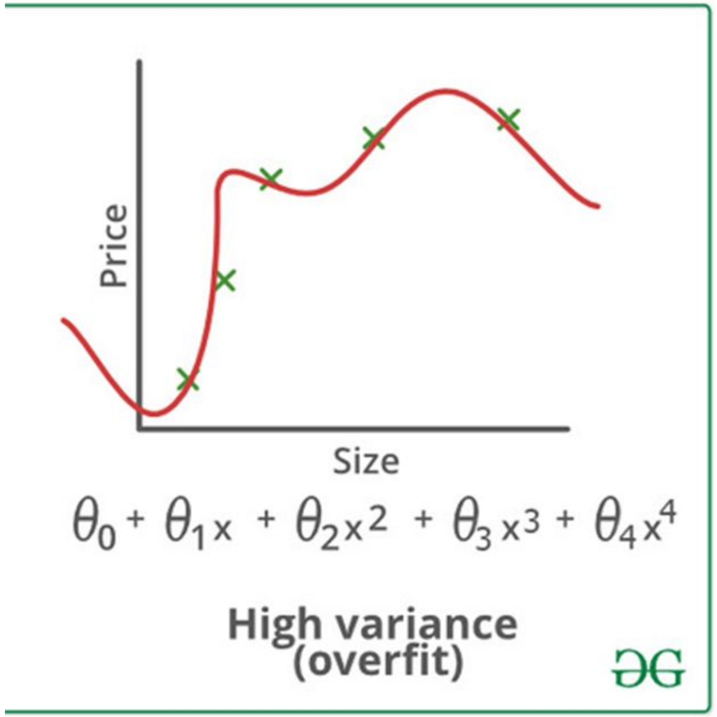


Underfitting

1. Reasons for Underfitting:
   a. High bias and low variance
   b. The size of the training dataset used is not enough.
   c. The model is too simple.
   d. Training data is not cleaned and also contains noise in it.
2. Techniques to reduce underfitting:
   a. Increase model complexity
   b. Increase the number of features, performing feature engineering
   c. Remove noise from the data.
   d. Increase the number of epochs or increase the duration of training to get better results.



Overfitting



1. Reasons for Overfitting are as follows:
    a. High variance and low bias
    b. The model is too complex
    c. The size of the training data
2. Techniques to reduce overfitting:
    a. Increase training data.
    b. Reduce model complexity.
    c. Early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
    d. Ridge Regularization and Lasso Regularization
    e. Use dropout for neural networks to tackle overfitting.

As Regularization

Regularization is a very important technique in machine learning to prevent overfitting. Mathematically speaking, it adds a regularization term in order to prevent the coefficients to fit so perfectly to overfit.

L1 Regularization

$$\text{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda \sum_{j=0}^{M} |W_j|$$

L2 Regularization

$$\text{Cost} = \underbrace{\sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2}_{\text{Loss function}} + \underbrace{\lambda \sum_{j=0}^{M} W_j^2}_{\text{Regularization Term}}$$

The difference between their properties can be promptly summarized as follows:

L2 regularization	L1 regularization
Computational efficient due to having analytical solutions	Computational inefficient on non-sparse cases
Non-sparse outputs	Sparse outputs
No feature selection	Built-in feature selection

Sparsity: some parameters become 0

Differences

The table below shows the summarized differences between L1 and L2 regularization

	L1 Regularization	**L2 Regularization**
1	Panelizes the sum of absolute value of weights.	penalizes the sum of square weights.
2	It has a sparse solution.	It has a non-sparse solution.
3	It gives multiple solutions.	It has only one solution.
4	Constructed in feature selection.	No feature selection.
5	Robust to outliers.	Not robust to outliers.
6	It generates simple and interpretable models.	It gives more accurate predictions when the output variable is the function of whole input variables.
7	Unable to learn complex data patterns.	Able to learn complex data patterns.
8	Computationally inefficient over non-sparse conditions.	Computationally efficient because of having analytical solutions.