Machine Learning

1. What is machine learning?

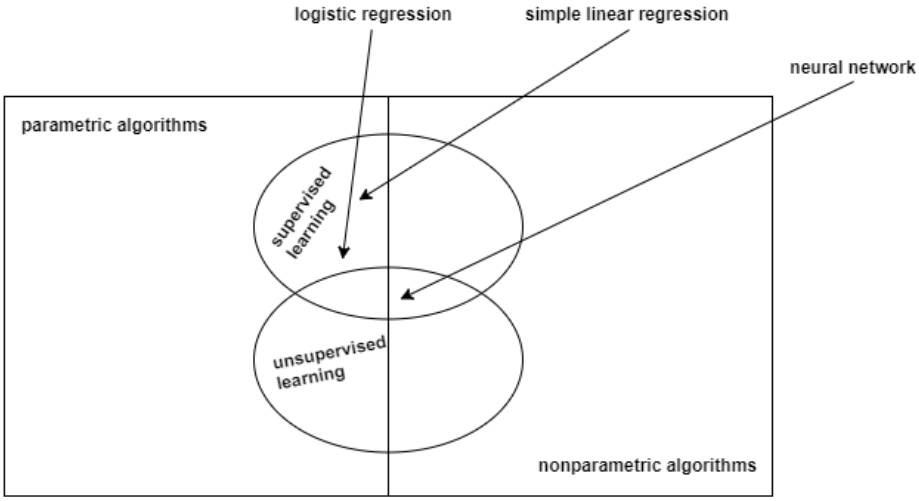
- A field of study that gives computers the ability to learn without being explicitly being programmed
- A computer is said to learn from experience E with respect to some tasks T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
- Machine learning algorithms are described as learning a **target function** (f) that best maps input variables (X) to an output variable (Y).

$$Y = f(X)$$

2. Machine learning algorithms:

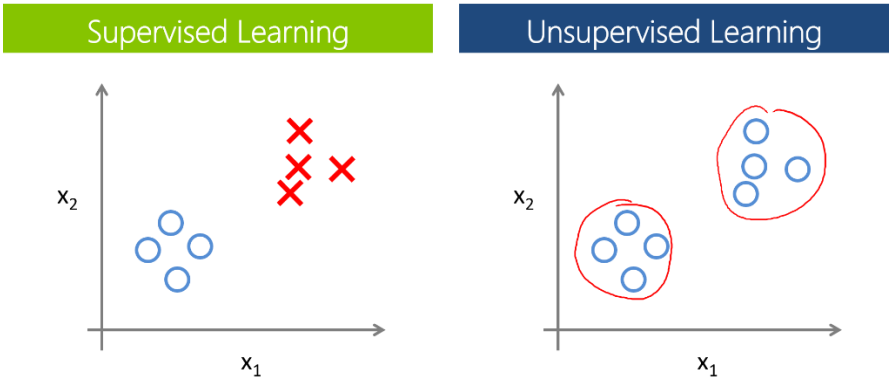
- **Parametric Machine Learning Algorithms**
  - Algorithms that simplify the function to a known form are called parametric machine learning algorithms.
  - The algorithms involve two steps:
    1. Select a form for the function.
    2. Learn the coefficients for the function from the training data.
  - Examples:
    - Logistic Regression
    - Linear Discriminant Analysis
    - Perceptron
- **Nonparametric Machine Learning Algorithms**
  - Algorithms that do not make strong assumptions about the form of the mapping function.
  - Examples:
    - Decision Trees like CART and C4.5
    - Naive Bayes
    - Support Vector Machines
    - Neural Networks
- Supervised learning
- Unsupervised learning
- Others: reinforcement learning, recommender systems

Machine Learning



3. Example of machine learning problems:

- Playing checkers
  - E = the experience of playing many games of checkers
  - T = the task of playing checkers
  - P = the probability that the program will win the next game



4. Supervised learning

- Example: housing price prediction
- In supervised learning, we are given a data set (training data) and already know what our correct output should look like, having the idea that there is a relationship between the input and output.
- Supervised learning problems are categorized into:
  - Regression problem:
    - Predict results within a continuous output (maps input variables to some continuous function)
    - Example: housing price prediction
  - Classification problem:

- Predict result in a discrete output (maps input variables into discrete categories)
- Example: determine whether a tumor is benign our malignant

5. Unsupervised learning
   - Idea: given a data set, can we find some structure within it?
   - Unsupervised learning is where you only have input data (X) and no corresponding output variables.
   - The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.
   - There is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data
   - Example:
     - Organize computer cluster
     - Social network analysis
     - Market segmentation
     - Astronomical data analysis
   - Unsupervised learning problems can be further grouped into clustering and association problems.
     - Clustering:
       - A clustering problem is where you want to discover the inherent groupings in the data
       - Example: grouping customers by purchasing behavior.
     - Association:
       - An association rule learning problem is where you want to discover rules that describe large portions of your data
       - Example: people that buy A also tend to buy B.
   - Unsupervised learning allows us to approach problems with little or no idea what our results should look like
   - We can derive structure from data where we don't necessarily know the effect of the variables
   - We can derive the structure by clustering the data based on relationships among the variables in the data
   - With unsupervised learning, there's no feedback based on the prediction results.
6. Bias-Variance Trade-Off
   - **Bias Error**
     - Bias are the simplifying assumptions made by a model to make the target function easier to learn.
     - Parametric algorithms have a high bias → fast to learn and easier to understand but less flexible.

        - Some example algorithms that have high bias include: Linear regression, logistic regression.
        - Example of low bias algorithms: Decision trees, support vector machine.
        - **Low Bias**: Suggest more assumptions about the form of the target function
        - **High-bias**: Suggest less assumptions about the form of the target function
    - **Variance Error**
        - Variance is the amount that the estimate of the target function will change if different training data was used.
        - **Low Variance**: Suggests small changes to the estimate of the target function with changes to the training dataset
        - **High Variance**: Suggests large changes to the estimate of the target function with changes to the training dataset.
        - Generally nonparametric machine learning algorithms that have a lot of flexibility have a high bias.
    - **Bias-Variance Trade-Off**
        - The goal of any supervised machine learning algorithm is to achieve low bias and low variance, however:
            - Increasing the bias will decrease the variance.
            - Increasing the variance will decrease the bias.
        - Parametric or linear machine learning algorithms often have a high bias but a low variance.
        - Nonparametric or nonlinear machine learning algorithms often have a low bias but a high variance.

7. Overfitting and Underfitting

- Poor performance in machine learning is either overfitting or underfitting the data.
- **Overfitting:**
    - Overfitting refers to a model that models the training data too well.
    - It has good performance on the training data but poor generalization to other data
    - The noise or random fluctuations in the training data is picked up and learned as concepts by the model.
    - Overfitting is more likely with nonparametric and nonlinear models that have more flexibility when learning a target function.
    - There are two techniques that you can use when evaluating machine learning algorithms to limit overfitting:
        - **Resampling technique to estimate model accuracy**

        - Using cross validation is a gold standard in applied machine learning for estimating model accuracy on unseen data.
        - Example: **k-fold cross validation**: allows you to train and test your model k-times on different subsets of training data and build up an estimate of the performance of a machine learning model on unseen data.
    - **Hold back a validation dataset**
        - A validation dataset is a subset of the training set that you hold back from your algorithm until the end of the project.
        - You can evaluate the learned models on the validation dataset to get a final objective idea of how the models might perform on unseen data.
- **Underfitting:**
    - Underfitting refers to a model that can neither model the training data nor generalize to new data.
    - An underfit machine learning model is not a suitable model and will have poor performance on the training data.
    - It has poor performance on the training data and poor generalization to other data