# Contents



## 0.1 Supervised Learning

Have a dataset with defined outputs that the system can be trained to predict.

## 0.2 Unsupervised Learning

Have a large amount of data with no defined labels. The model should be able to identify similar groups even if those groups aren't predefined.

## 0.3 Reinforcement Learning

The agent that needs to learn, will generate its own data through freely interacting with the environment. Through its own experience, it will generate data which it can learn off of.

## 0.4 Transfer Learning

A model trained off of dataset $D_1$ can be retrained with dataset $D_2$ to add more knowledge to the model. This means that a model does not have to be trained from scratch.

## 0.5 Performance Measure, $P$

**Accuracy:** The proportion of examples for which the model produces the correct output.

**Error rate:** The proportion of examples for which the model produces an incorrect output.

**Loss function:** Quantifies the difference between the predicted outputs of a machine learning algorithm and the actual target values.

**Generalization** Ability to perform well on previously unobserved data; e.g., evaluate the performance using a test set.

## 0.6 No Free Lunch Theorem

- "No Free Lunch Theorem": without having substantive information about the modeling problem, there is no single model that will always do better than any other model.
- The goal of machine learning research is not to seek a universal learning algorithm or the absolute best learning algorithm.
- Instead, out goal is to understand what kinds of distributions are relevant to the real world and what kinds of machine learning algorithms perform well on data drawn distributions we care about.

## 0.7 Training and Testing

- Training data: used to train the machine learning model.
- Testing data: used to determine the performance of the trained model.

### 0.7.1 Independent and Identically Distributed (IID) assumptions

- Examples in each dataset are independent from each other
- Training and testing set are identically distributed; i.e., drawn from the same probability distribution as each other.

## 0.8 Fitting

- Underfitting: when the model is unable to obtain a sufficiently low training value.
- Overfitting: when the gap between the training error and test error is too large.
- Hypothesis: the machine's presumption regarding the connection between the input features and the output.

Consider a hypothesis $h$ and its...

- Error rate over training data: $error_{train}(h)$
- True error rate over all data: $error_{true}(h)$

Hypothesis $h$ overfits the training error if there is an alternative hypothesis $h'$ such that

$$\begin{aligned} error_{train}(h) &< error_{train}(h') \\ error_{true}(h) &> error_{true}(h') \end{aligned} \tag{1}$$

### 0.8.1 Resolving Underfitting

- Increasing model complexity.
- Using a different ML algorithm.
- Increasing the amount of training data.
- Ensemble methods to combine multiple methods to create better outputs.
- Feature engineering for creating new model features from the existing ones that may be more relevant to the learning task.

### 0.8.2 Resolving Overfitting

- Cross-Validation: a technique for evaluating ML models by training several ML models on subsets of the available input data and evaluating them on another subset of the data.
- Regularization: a technique where a penalty term is added to the loss function, discouraging the model from assigning too much importance to individual features.
- Early-stopping: stops training when a monitored metric has stopped improving.
- Bagging: learning multiple models in parallel and applying majority voting to choose the final candidate model.

## 0.9 Cross-Validation

### 0.9.1 $k$-fold cross-validation

- Divide data into $k$ folds.
- Train on $k-1$ folds, use the $k^{\text{th}}$ to measure error.
- Repeat $k$ times; use average error to measure generalization accuracy.
- Statistically valid and gives good accuracy estimates.

### 0.9.2 Leave-one-out cross validation (LOOCV)

## 0.10 Parametric Learning

Parametric learning algorithms: make strong assumptions about the form of the mapping function between the input features and output.

- For example, logistic regression, linear regression, perceptron, naïve bayes, neural network.
- Benefits of such models are
  (a) easier to understand and interpret results
  (b) very fast to learn from data
  (c) do not require as much training data
  (d) can work even if they do not fit the data perfectly.
- However, by pre-emptively choosing a functional form, these methods are highly constrained to the specified form.

## 0.11 Non-parametric Learning

Non-parametric learning algorithms: do not make assumptions about the form of the mapping function between the input features and output.

- For example, SVM, $k$-NN, $k$-means, decision tree.
- Benefits include
  (a) being capable of fitting a large number of functional forms,
  (b) there are no assumptions about the underlying....

## 0.12 Model Selection

- Adopting the best algorithm and model for a specific dataset by assessing and comparing different models to identify the one with the best results.

## 0.13 AIC/BIC

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): compares different models to choose one that best fits the data.

- The goal of both AIC and BIC is to balance the goodness-of-fit of the model with its complexity, in order to avoid overfitting or underfitting.
- Both AIC and BIC penalize models with large number of parameters relative to the size of the data, but BIC penalizes more severely.

$$\begin{aligned} \min AIC &= \min\{2m - 2\log(L)\} \\ \min BIC &= \min\{m\log(n) - 2\log(L)\} \end{aligned} \tag{2}$$

where $m$ is the number of model parameters, $n$ is the number of data points, and $L$ is the maximum likelihood of the model.

## 0.14 Classification