Machine Learning Concepts

ML attempts to learn models of the world; many types of data are used to train these models. The data defines the form of learning that can be used.

Unsupervised Learning:

Only the input data is provided and the models learn to extract patterns from data.

- Clustering - Grouping similar items together
- Dimensionality reduction
- Density Estimation - Estimating probabilities of input data points

Supervised Learning:

Each input is paired with an output label and the model is trained to minimise the error between its output and the label assigned to that data point.

- Regression: numerical outputs
- Classification: category labels

Reinforcement Learning:

The correct outputs are not given but the model receives a reward or punishment depending on its actions. It deals with dynamic environments where agents carry out sequences of actions.

Model Selection:

Train the model on 80% of the dataset and use the other 20% for validating the model.

- Cross-validation is the concept of splitting the dataset into S groups so that (S - 1) / S data is used from training. For example if S = 4, we have 4-fold cross-validation using 3 subsets for training and 1 for validation. The model is trained independently on the combinations of the 3 subsets.

Dimensionality:

Computing models on high-dimensionality datasets can be infeasible. However real-world data tends to be restricted to regions of the much bigger spaces. The effective dimensionality is often smaller than the space.