Introduction to Machine Learning

**Glossary**

**Bag of words**: A technique used to extract features from the text. It counts how many times a word appears in a document (corpus), and then transforms that information into a dataset.

A **categorical**label has a discrete set of possible values, such as "is a cat" and "is not a cat."

**Clustering**. Unsupervised learning task that helps to determine if there are any naturally occurring groupings in the data.

**CNN**: Convolutional Neural Networks (CNN) represent nested filters over grid-organized data. They are by far the most commonly used type of model when processing images.

A **continuous (regression)** label does not have a discrete set of possible values, which means possibly an unlimited number of possibilities.

**Data vectorization**: A process that converts non-numeric data into a numerical format so that it can be used by a machine learning model.

**Discrete**: A term taken from statistics referring to an outcome taking on only a finite number of values (such as days of the week).

**FFNN**: The most straightforward way of structuring a neural network, the Feed Forward Neural Network (FFNN) structures neurons in a series of layers, with each neuron in a layer containing weights to all neurons in the previous layer.

**Hyperparameters**are settings on the model which are not changed during training but can affect how quickly or how reliably the model trains, such as the number of clusters the model should identify.

**Log loss** is used to calculate how uncertain your model is about the predictions it is generating.

**Hyperplane**: A mathematical term for a surface that contains more than two planes.

**Impute**is a common term referring to different statistical tools which can be used to calculate missing values from your dataset.

**label**refers to data that already contains the solution.

**loss function** is used to codify the model’s distance from this goal

**Machine learning**, or ML, is a modern software development technique that enables computers to solve problems by using examples of real-world data.

**Model accuracy**is the fraction of predictions a model gets right. Discrete: A term taken from statistics referring to an outcome taking on only a finite number of values (such as days of the week). Continuous: Floating-point values with an infinite range of possible values. The opposite of categorical or discrete values, which take on a limited number of possible values.

**Model inference** is when the trained model is used to generate predictions.

model is an extremely generic program, made specific by the data used to train it.

**Model parameters**are settings or configurations the training algorithm can update to change how the model behaves.

**Model training algorithms** work through an interactive process where the current model iteration is analyzed to determine what changes can be made to get closer to the goal. Those changes are made and the iteration continues until the model is evaluated to meet the goals.

**Neural networks**: a collection of very simple models connected together. These simple models are called **neurons**. The connections between these models are trainable model parameters called **weights**.

**Outliers**are data points that are significantly different from others in the same sample.

**Plane**: A mathematical term for a flat surface (like a piece of paper) on which two points can be joined by a straight line.

**Regression**: A common task in supervised machine learning.

In**reinforcement learning**, the algorithm figures out which actions to take in a situation to maximize a reward (in the form of a number) on the way to reaching a specific goal.

**RNN/LSTM**: Recurrent Neural Networks (RNN) and the related Long Short-Term Memory (LSTM) model types are structured to effectively represent for loops in traditional computing, collecting state while iterating over some object. They can be used for processing sequences of data.

**Silhouette coefficien**t: A score from -1 to 1 describing the clusters found during modeling. A score near zero indicates overlapping clusters, and scores less than zero indicate data points assigned to incorrect clusters. A

**Stop words**: A list of words removed by natural language processing tools when building your dataset. There is no single universal list of stop words used by all-natural language processing tools.

In **supervised learning,** every training sample from the dataset has a corresponding label or output value associated with it. As a result, the algorithm learns to predict labels or output values.

**Test dataset**: The data withheld from the model during training, which is used to test how well your model will generalize to new data.

**Training dataset**: The data on which the model will be trained. Most of your data will be here.

**Transformer**: A more modern replacement for RNN/LSTMs, the transformer architecture enables training over larger datasets involving sequences of data.

In **unlabeled data**, you don't need to provide the model with any kind of label or solution while the model is being trained.

In **unsupervised learning**, there are no labels for the training data. A machine learning algorithm tries to learn the underlying patterns or distributions that govern the data.