University of Essex
MSc Artificial Intelligence
Module: UAI_PCOM7E January 2025 A
Unit 6: Collaborative Discussion – Supervised v. Unsupervised Machine Learning Algorithms

Initial Discussion

Two fundamental algorithms used in supervised machine learning are linear and logistic regression, while supervised machine learning itself, in general, refers to using sets of known and pre-categorized data points to train predictive models to find and correctly identify or categorize unknown data points consistent with the boundaries of a model (Burkov, 2019).

Specifically, linear regression models attempt to predict new values from new inputs consistent with a linear, one-to-one relationship established from a training data set. For example, a linear regression model could be used to predict next month's average rainfall given an input of the current month's accumulated rainfall. Linear regression does so by minimizing the overall chance of a prediction being incorrect, or the aggregate "wrongness" of a prediction for any given input value (Burkov, 2019).

An advantage of linear regression is its relative simplicity to implement (Burkov, 2019). A recurring pitfall is that the relationship to be defined might not stay factually consistent with a true one-to-one correspondence between input values and output predictions; another risk is overfitting, or that the training data set may accurately describe a linear relationship within its own boundaries, but fails to accurately conform new input values to valid predicted values (Hewamalage, Ackermann, and Bergmeir, 2022).

Logistic regression is typically used for either-or, binary categorization tasks, such as identifying whether a new input does or does not belong, or how likely the input does or does not belong given a defined margin of error, within a category established by the training data set (Burkov, 2019). For example, in image identification tasks, the categorization could involve identifying whether or not a given image is mostly likely to be categorized as "a car" or "not a car."

An advantage of logistic regression, shared with linear regression, is that it is relatively straightforward to create a model (Burkov, 2019). A common risk is that the scope features of the category defined by the dataset aren't truly the features needed to classify new inputs (Hewamalage, Ackermann, and Bergmeir, 2022). In image classification, a model that doesn't identify the correct set of features to distinguish between accurate and inaccurate classifications might attempt to classify

a cardboard box as “a car” given the blocky shape, while missing other relevant classification features (such as that cars generally have four wheels).

References

Burkov, A. (2019). *The Hundred-Page Machine Learning Book*. Andriy Burkov.

Hewamalage, H., Ackermann, K. and Bergmeir, C. (2022). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*. doi:https://doi.org/10.1007/s10618-022-00894-5.