# HOME LEARNING TASK

## Logistic Regression Summary

This machine learning algorithm is an artificial intelligence approach that programmers use to predict how an observation, in best probability, will be classified.

It is able to predict and classify data through semi-supervised learning: the model is given past data and the labels of which binary class the data belongs to, so that the ML algorithm can identify those classes in the future.

The algorithm takes input variables and produces expected outputs. Depending on the desired predicted value data scientists can go back and asses which input variables cause the greatest shift in prediction outputs. Logistic Regression is also a statistical model using the equation:

$$y = \frac{1}{1+e^{-(w_o + w_1 x)}}$$

Where two classification outcomes (y) are given on a scale of 0 to 1. When input data is first given to the ML algorithm y is the target variable, the output that the algorithm is trying to predict. In order to classify the predicted observations programmers (and statisticians) use a decision boundary such as

*y=0 if the predicted probability is <0.5*

*y=0 if the predicted probability is >0.5*

This would be typically seen in binary logistic regression but there are other types of logistic regressions too such as multinomial logistic regression and ordinal logistic regression, where y has three or more possible classes, A or B or C:

*y with target class A= the predicted probability of A vs B and C*

*y with target class B= the predicted probability of B vs A and C*

*y with target class C= the predicted probability of C vs A and B*

(And this can be repeated depending on the number of classes).

Logistic Regression is great because it can be applied on a range of scales, when given enough input data to train with, and training is not overly complicated for software engineers because it is not a deep learning model that uses neural networks. The ML algorithm can continue to train and produce more accurate outcomes near real-time. Logistic Regression is most useful when predicting medical radiological images, bank loan

repayments, weather and storm predictions, qualifying business leads and customers, upselling products digitally to specific customers.

A good Logistic Regression will have accuracy, hopefully more than 95% of the data will be correctly predicted and classified, and a good Logistic Regression will have ROC AUC which stands for Area Under the Receiver Operating Characteristic Curve and is just a ratio of the correctly predicted data vs the incorrectly predicted data.

One example of a Logistic Regression algorithm is used by the company Microculus, which produce blood testing kits. Using the large amounts of available medical research on molecules and compounds they worked with Microsoft to create Loom: a Logistic Regression ML algorithm to identify genetic diseases. They could have used Support Vector Machine algorithms or Random Forest, but Logistic Regression produced the quickest and most accurate results. Loom is a text analyser, trained with labelled input data that extracts key phrases or entities from scientific research papers, to predict whether there's a link between microRNA and certain genes in blood tests, by giving a score for the relationship between the microRNA and the specific gene being studied.