Machine Learning Task

Bronwyn Bowles-King

1. For each of the following examples describe at least one possible input and output. Justify your answers:

1.1. A self-driving car

Considering that "input is whatever we want the algorithm to learn from", video of surrounding roads, sensors for nearby objects, a satellite navigation system, and traffic updates would all be suitable inputs for a self-driving car (HyperionDev, 2025, p. 3). As "output is the outcome we want the algorithm to be able to produce", instructions to the vehicle systems, such as the gears, brakes and clutch, to drive forward, slow down, turn, etc., in a safe manner would be the desired output (ibid).

1.2. Netflix recommendation system

The input would be the user's search, viewing and browsing history, especially for shows and films that they finished watching, rewatched or saved on their profile. The input may also include shows and films that similar users watched. The output would be new shows and films that have been placed in genres that the user showed a preference for.

1.3. Signature recognition

The input would be a large set of handwritten signatures that trained the program on which examples are similar enough to be accepted as being by the same person. Forged signatures could also be included to teach the program what to detect as problematic in a signature. The output would be correctly identifying a scanned signature or one made with a touchscreen as a signature first, and then as the same as the original reference saved.

1.4. Medical diagnosis

The input would include a list of the person's symptoms, their vital signs, scans of their body, such as x-rays or pictures of their skin if it is a dermal disease, test results, such as blood or urine tests, and other people's observations, such as by parents or nurses. The output would be a correct diagnosis for the person. In some cases, it may be the suggestion of various possibilities if it is not possible to make a firm diagnosis at the time, and that more tests are needed to determine the cause.

2. For each of the following case studies, determine whether it is appropriate to utilise regression or classification machine learning algorithms. Justify your answers:

2.1. Classifying emails as promotional or social based on their content and metadata.

> Classification machine learning (ML) is appropriate here as the emails must be grouped into categories by features that are not generally numerical.

2.2. Forecasting the stock price of a company based on historical data and market trends.

> Regression ML is suited as the algorithm will work with numerical data and possibly time periods to make estimates and best predictions.

2.3. Sorting images of animals into different species based on their visual features.

> Classification ML is appropriate here as the images should be categorised by their physical features.

2.4. Predicting the likelihood of a patient having a particular disease based on medical history and diagnostic test results.

> A medical diagnosis program would need to be quite advanced and handle different input types. It may work best to combine regression and classification ML, as the inputs can vary from categorical-type data on symptoms to numerical data from test results.

3. For each of the following real-world problems, determine whether it is appropriate to utilise a supervised or unsupervised machine learning algorithm. Justify your answers:

3.1. Detecting anomalies in a manufacturing process using sensor data without prior knowledge of specific anomaly patterns.

> As the algorithm is not given previous data or training to work from, this is unsupervised ML.

3.2. Predicting customer lifetime value based on historical transaction data and customer demographics.

> The algorithm has historical and customer data to compare before making a prediction and thus, this is supervised ML.

3.3. Segmenting customer demographics based on their purchase history, browsing behaviour, and preferences.

The program is given data on browsing and purchasing history and personal preferences before placing customers in a demographic group, so this may be supervised ML. However, it is not clear whether the provided data is also linked to demographics, and so this can be a semi-supervised approach here.

3.4. Analysing social media posts to categorise them into different themes.

This is unsupervised ML because it is not stated that the computer has training data or labels upon which to categorise the data by theme.

4. For each of the following real-world problems, determine whether it is appropriate or inappropriate to utilise semi-supervised machine learning algorithms. Justify your answers:

4.1. Predicting fraudulent financial transactions using a dataset where most transactions are labelled as fraudulent or legitimate.

Semi-supervised learning uses labelled and unlabelled data together, but as most of the dataset is labelled in this case, this is more of a supervised approach. Greater supervision seems necessary, considering the sensitivity of financial transactions being monitored and the risks involved.

4.2. Analysing customer satisfaction surveys where only a small portion of the data is labelled with satisfaction ratings.

Semi-supervised learning uses a small set of labelled data generally, and so it is appropriate in this case. If well designed, the program will likely be able to learn quite well how to analyse a customer survey with less supervision, especially from a large survey and if the survey gathers mostly numerical data (e.g. Likert-scale user feedback).

4.3. Identifying spam emails in a dataset where the majority of emails are labelled.

Most of the emails are already labelled, so this is supervised ML. Much data may be available because users have already indicated which emails they consider spam, giving the program plenty to work with.

4.4. Predicting the probability of default for credit card applicants based on their complete financial and credit-related information.

The large amount of data available on applicants' financial history can be semi-supervised ML if some of it is labelled. If it is not labelled, then unsupervised ML would be more appropriate. However, considering the sensitivity of personal financial data and risks involved, a supervised approach seems best in this case.

References

Delua, J. (2021). Supervised versus unsupervised learning: What's the difference? https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning

HyperionDev. (2025). Machine Learning. Course materials. Private repository, GitHub.

HyperionDev. (2025). Data Visualisation – Tableau. Course materials. Private repository, GitHub.