Introduction to Machine Learning

Martha Lewis

Department of Engineering Mathematics
University of Bristol
martha.lewis@bristol.ac.uk

Outline

- **Types of learning problem**
- **Classification vs regression**
- Performance metrics
- Generalisation and overfitting
- Train, validate, test

Types of Learning Problem

- Supervised learning
- Unsupervised learning
- (Reinforcement learning)

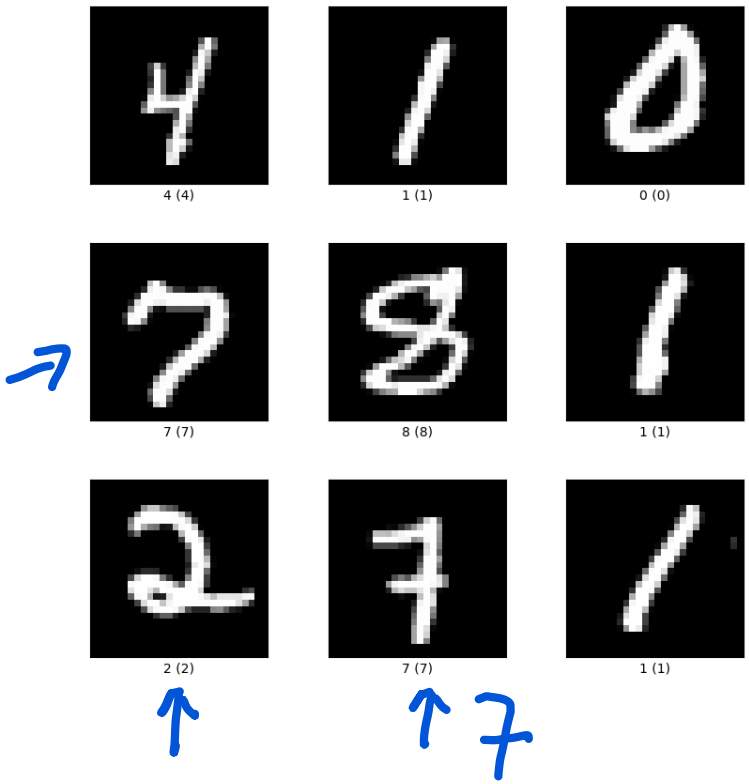
Types of Learning Problem

Supervised Learning: Learning an unknown function $f : \Omega \rightarrow \mathcal{O}$ where Ω is an n-dimensional feature space and $\mathcal{O}$ is an output space (typically 1-dimensional).

- The aim is to learn an approximation of f from a set of labelled patterns of the form $(x_1, \ldots, x_n, y)$ where $y = f(x_1, \ldots, x_n)$.

Supervised Learning Examples

MNIST image classification



Feature space: each image is 28 by 28 (=784) pixels, so each image could be represented by a 784-length vector $\vec{x} \in \mathbb{R}^{784}$

Output space: The output space is a number in the set $\{0, 1, 2, ..., 9\}$

Supervised Learning Examples

Sentiment analysis

- A monumental big-screen spectacle - POSITIVE
- Sweeps you away on waves of pure, ravishing cinema - POSITIVE
- The plot is rote and the furnishings tired - NEGATIVE
- For all the nifty bits of reverse chronology, there's little that lingers in the imagination - ?

What is the input space? Could be: the set of vocabulary in the sentences, or a vector space where we can represent sentences as vectors, ...
What is the output space? Could be the set $\{\text{POS, NEG, NEUTRAL}\}$, or $[0, 1]$, other options too.

Supervised Learning Examples

House Prices

Crime	Industry	Rooms	Tax	Value
0.006	2.31	6.6	296	24.0
0.027	7.07	6.4	242	21.6
0.027	7.07	7.2	242	34.7
0.032	2.18	7.0	222	33.4
0.069	2.18	7.1	222	?

Feature space: could be $\mathbb{R}^4$

Output space: could be $\mathbb{R}$

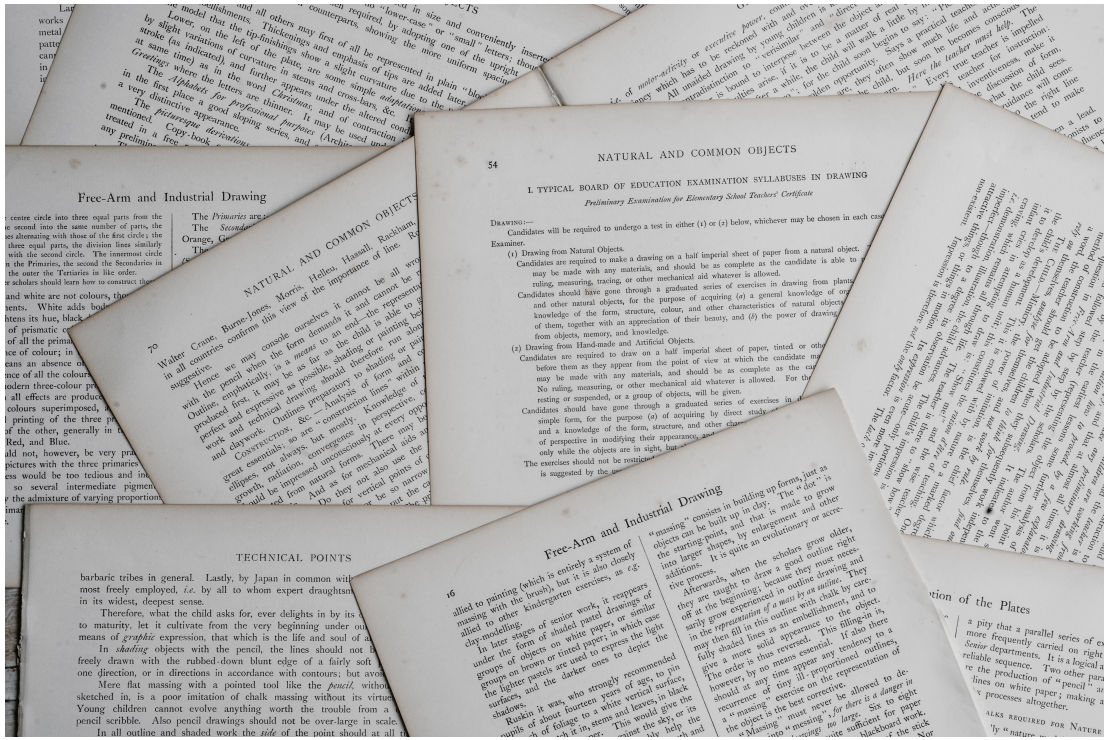
Types of Learning Problem

Unsupervised Learning: Identify (unknown) patterns in the feature space Ω based on unlabelled data.

- Notions of similarity or distance between elements of the feature space play a predominant role.

Unsupervised Learning Examples

Information Retrieval



Clustering algorithms identify groups of similar documents. Given a new document, we can find others that are similar by calculating the distance in the feature space.

Feature space could be the set of words in the vocabulary, or a vector space in which documents can be represented.

Unsupervised Learning Examples

Fraud detection

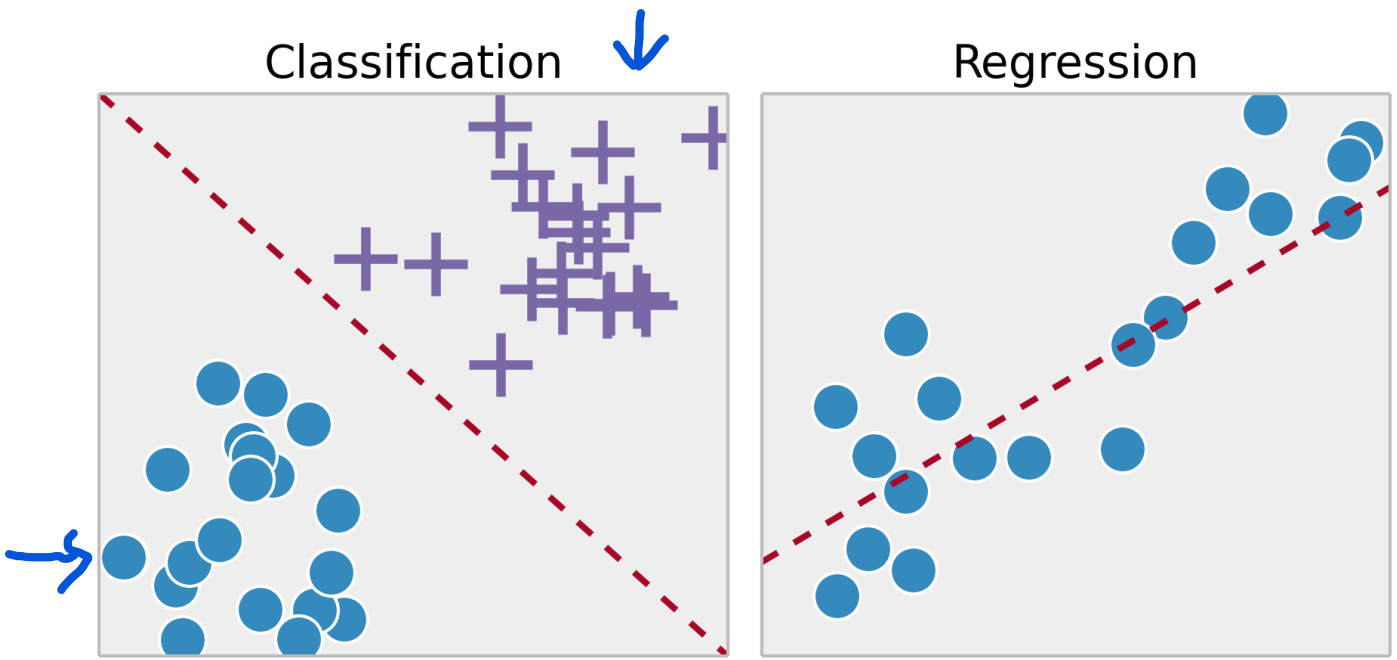
Distance in the feature space is used to detect when a transaction is fraudulent. Legitimate transactions form clusters in the feature space. Transactions outside of these clusters are flagged for fraud. Feature space could include location, vendor, time of day...

`https://www.vecteezy.com/free-vector/fraud`

Classification vs Regression

In supervised learning, the problem is to learn an unknown function $f : \Omega \rightarrow \mathcal{O}$ where Ω is an n-dimensional feature space and $\mathcal{O}$ is an output space (typically 1-dimensional).

- A *classification problem* is where $\mathcal{O} = \{c_1, \ldots, c_k\}$, corresponding to a finite set of classes or categories.
- A *regression problem* is where $\mathcal{O} = \mathbb{R}$.



Classification vs. Regression: Examples

Are the following classification or regression problems?

- MNIST image classification task
- Sentiment analysis
- House price prediction

Classification vs. Regression: Examples

Are the following classification or regression problems?

- MNIST image classification task
- Sentiment analysis
- House price prediction

Pause the video now ...

Classification vs. Regression: Examples

Are the following classification or regression problems?

- MNIST image classification task
  Classification - output space is $\{0, 1, ..., 9\}$
- Sentiment analysis
  Classification - output is {POS, NEG, NEUTRAL} or
  Regression - output is $[0, 1]$
- House price prediction
  Regression - output space is (some subset of) $\mathbb{R}$