Machine Learning Concepts

Prof. Dr. H. H. Takada

Quantitative Research – Itaú Asset Management
Institute of Mathematics and Statistics – University of São Paulo

What is Machine Learning?

Field of study that develop methods, techniques, models and algorithms with the ability to learn.

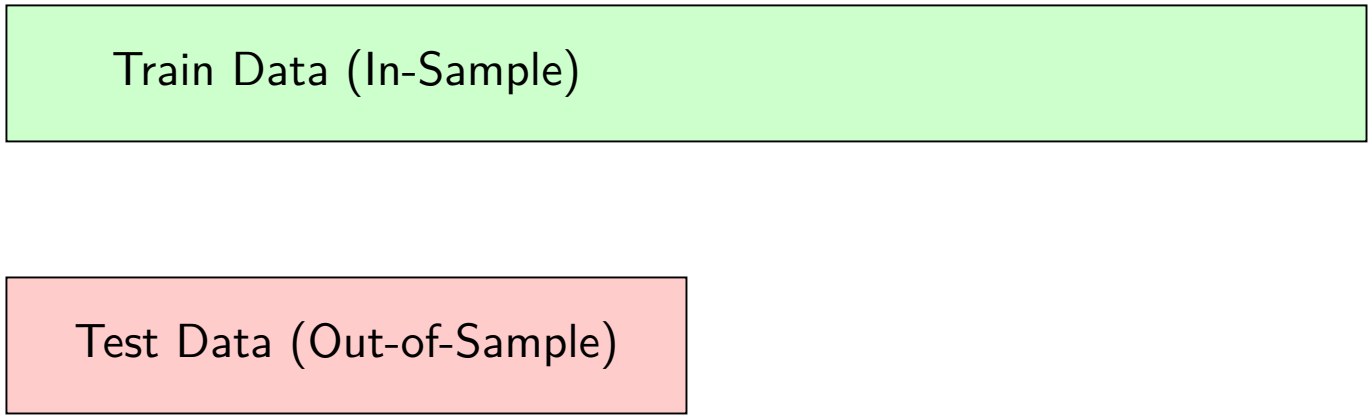
Supervised and Unsupervised Learning

Broadly speaking the main two subfields of Machine Learning are *supervised learning* ('input' and 'output' data available) and *unsupervised learning* (only 'inputs' available).

- In supervised learning the focus is on accurate prediction.
- In unsupervised learning the aim is to find compact descriptions of the data.

In both cases, one is interested in methods that generalize well to previously unseen data. In this sense, one distinguishes between data that is used to train a model and data that is used to test the performance of the trained model.

Train versus Test data



- The parameters of the model are set on the basis of the train data only.
- If the test data is generated from the same underlying process that generated the train data, an unbiased estimate of the performance can be obtained by measuring the test data performance of the trained model.
- Importantly, the test performance should not be used to adjust the model parameters since we would then no longer have an independent measure of the performance of the model.

Supervised learning example

- Consider a database of face images, each represented by a vector $\mathbf{x}$.
- Along with each image $\mathbf{x}$ is an output class $y \in \{\text{male}, \text{female}\}$ that states if the image is of a male or female.
- A database of 10,000 such image-class pairs is available, $\mathcal{D} = \{(\mathbf{x}^n, y^n), n = 1, \ldots, 10000\}$.
- The task is to make an accurate predictor $y(\mathbf{x}^*)$ of the sex of a novel image $\mathbf{x}^*$.
- This is an example application that would be hard to program in a traditional manner since formally specifying a rule that differentiates male from female faces is difficult. An alternative is to give example faces and their gender labels and let a machine automatically 'learn' such a rule.

Supervised Learning

- Given a set of data $\mathcal{D} = \{(x^n, y^n), n = 1, \ldots, N\}$ the task is to learn the relationship between the input x and output y such that, when given a novel input x^* the predicted output y^* is accurate.
- The pair (x^*, y^*) is not in $\mathcal{D}$ but assumed to be generated by the same unknown process that generated $\mathcal{D}$.
- To specify explicitly what accuracy means one defines a loss function $L(y^{pred}, y^{true})$ or, conversely, a utility function $U = -L$.
- Our interest is to describe y conditioned on x, i. e. $p(y|x, \mathcal{D})$. 'supervised' indicates that there is a notional 'supervisor' specifying the output y for each input x in the available data $\mathcal{D}$. The output is also called a 'label', particularly when discussing classification.
- Predicting tomorrow's stock price $y(T+1)$ based on past observations $y(1), \ldots, y(T)$ is a form of supervised learning. We have a collection of times and prices $\mathcal{D} = \{(t, y(t)), t = 1, \ldots, T\}$ where time t is the input and the price $y(t)$ is the output.

Classification and Regression

- If the output is one of a discrete number of possible 'classes', this is called a *classification* problem. In classification problems we will generally use c for the output.
- If the output is continuous, this is called a *regression problem*. For example, based on historical information of demand for sun-cream in your supermarket, you are asked to predict the demand for the next month.
- In some cases it is possible to discretize a continuous output and then consider a corresponding classification problem.
- However, in other cases it is impractical or unnatural to do this, for example if the output y is a high dimensional continuous valued vector, or if the ordering of states of the variable is meaningful.

Example

- A father decides to teach his young son what a sports car is. Finding it difficult to explain in words, he decides to give some examples. They stand on a motorway bridge and, as each car passes underneath, the father cries out 'that's a sports car!' when a sports car passes by. After ten minutes, the father asks his son if he's understood what a sports car is. The son says, 'sure, it's easy'.
- An old red VW Beetle passes by, and the son shouts – 'that's a sports car!'. Dejected, the father asks – 'why do you say that?'. 'Because all sports cars are red!', replies the son.

Example

- Here the father plays the role of the supervisor, and his son is the 'student' (or 'learner'). It's indicative of the kinds of problems encountered in Machine Learning in that it is not easy to formally specify what a sports car is – if we knew that, then we wouldn't need to go through the process of learning.
- This example also highlights the issue that there is a difference between performing well on training data and performing well on novel test data. The main interest in supervised learning is to discover an underlying rule that will generalize well, leading to accurate prediction on new inputs.
- If there is insufficient train data then, as in this scenario, performance may be disappointing.

Unsupervised learning

- Given a set of data $\mathcal{D} = \{x^n, n = 1, \ldots, N\}$ in unsupervised learning we aim to find a plausible compact description of the data.
- An objective is used to quantify the accuracy of the description. In unsupervised learning there is no special prediction variable so that, from a probabilistic perspective, we are interested in modelling the distribution $p(x)$. The likelihood of the model to generate the data is a popular measure of the accuracy of the description.

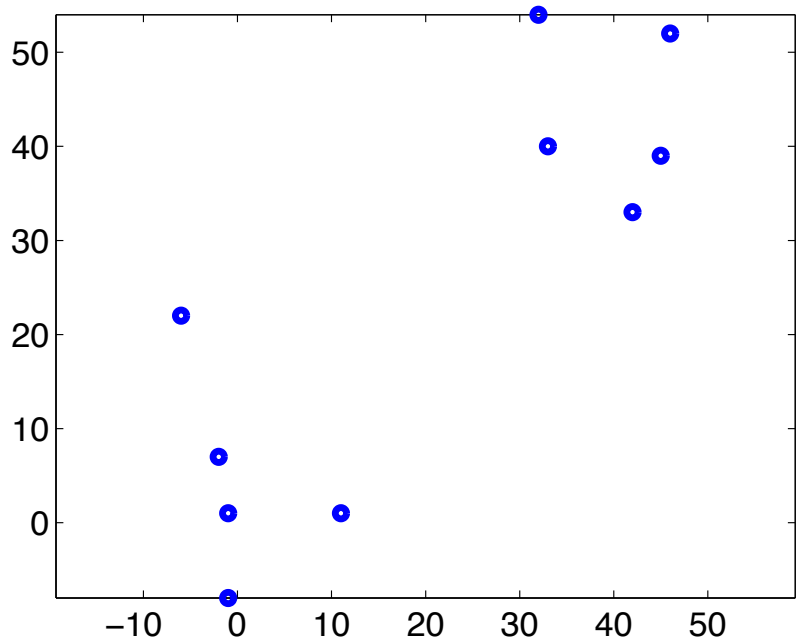
Example

coffee	1	0	0	1	0	0	0	$\cdots$
tea	0	0	1	0	0	0	0	$\cdots$
milk	1	0	1	1	0	1	1	$\cdots$
beer	0	0	0	1	1	0	1	$\cdots$
diapers	0	0	1	0	1	0	1	$\cdots$
aspirin	0	1	1	0	1	0	1	$\cdots$

- Each column represents the products bought by a customer (7 customer records with the first 6 of the 10,000 products are shown). A 1 indicates that the customer bought that item. We wish to find common patterns in the data, such as if someone buys diapers they are also likely to buy aspirin.
- A supermarket chain wishes to discover how many different basic consumer buying behaviours there are based on a large database of supermarket checkout data.
- Items brought by a customer on a visit to a checkout are represented by a (very sparse) 10,000 dimensional vector $\mathbf{x}$ which contains a 1 in the i^{th} element if the customer bought product i and 0 otherwise.
- Based on 10 million such checkout vectors from stores across the country, $\mathcal{D} = \left\{\mathbf{x}^n, n = 1, \ldots, 10^7\right\}$ the supermarket chain wishes to discover patterns of buying behaviour.

Clustering

x_1	-2	-6	-1	11	-1	46	33	42	32	45
x_2	7	22	1	1	-8	52	40	33	54	39



The table represents a collection of unlabelled two-dimensional points. A reasonable compact description of the data is that is has two clusters, one centered at (0,0) and one at (35,45), each with a standard deviation of 10.

Anomaly detection

- A baby processes a mass of initially confusing sensory data. After a while the baby begins to understand her environment so that sensory data from the same environment is familiar or expected.
- When a strange face presents itself, the baby recognizes that this is not familiar and becomes upset. The baby has learned a representation of the environment and can distinguish the expected from the unexpected; this is an example of unsupervised learning.
- Detecting anomalous events in industrial processes (plant monitoring), engine monitoring and unexpected buying behaviour patterns in customers all fall under the area of anomaly detection. This is also known as 'novelty' detection.

Online (sequential) learning

- In the above situations we assumed that the data was given beforehand.
- In *online learning* data arrives sequentially and we continually update our model as new data becomes available.
- Online learning may occur in either a supervised or unsupervised context.

Interacting with the environment

In certain situations, an agent may be able to interact in some manner with its environment. This interaction can complicate but also enrich the potential for learning.

Query (Active) Learning

Here the agent has the ability to request data from the environment. For example, a predictor might recognize that it is less confidently able to predict in certain regions of the space x and therefore requests more training data in this region. Active learning can also be considered in an unsupervised context in which the agent might request information in regions where $p(x)$ is currently uninformative.

Reinforcement Learning

In reinforcement learning an agent inhabits an environment in which it may take actions. Some actions may eventually be beneficial (lead to food for example), whilst others may be disastrous (lead to being eaten for example). Based on accumulated experience, the agent needs to learn which action to take in a given situation in order to maximize the probability of obtaining a desired long term goal (long term survival, for example). Actions that lead to long term rewards need to be reinforced. Reinforcement learning has connections with control theory, Markov decision processes and game theory.

Semi-supervised learning

- In Machine Learning, a common scenario is to have a small amount of labelled and a large amount of unlabelled data.
- For example, it may be that we have access to many images of faces; however, only a small number of them may have been labelled as instances of known faces.
- In semi-supervised learning, one tries to use the unlabelled data to make a better classifier than that based on the labelled data alone. This is a common issue in many examples since often gathering unlabelled data is cheap (taking photographs, for example).
- However, typically the labels are assigned by humans, which is expensive.
- Obtaining supervised training labels can also be accomplished using mechanisms such as Mechanical Turk (a crowdsourcing Internet marketplace enabling individuals and businesses, known as requesters, to coordinate the use of human intelligence to perform tasks that computers are currently unable to do).