5 Three components of ML

Date	@October 24, 2024
Topic	Theory

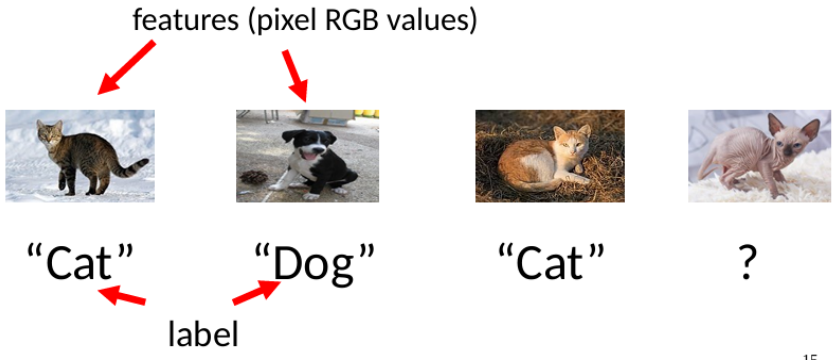
The big three components of ML are:

- data
- model
- loss

Data

Data is a set or a collection of data points. Data points are objects, record, cases, samples, entities or instances. They carry information which are referred to as features. A data point could be a person, an image, a signal, a server, and so on.

The features are low-level properties and are often easy to measure or compute. the labels are high-level quantity of interest and are often difficult to measure or compute.



We mainly use numeric features.

Label

Label is design choice, it means you choose what to consider as label of a data point. It is also called output variable, target or response variable. By choosing label you define the ML problem or learning task.

Label can be;

- categorical: it makes the problem a classification task; if there are only 2 categories it is a binary classification, otherwise it is a multi-class classification
- numerical: it makes the problem a regression task; if a data point have more than one label it is called a multi-label problem, otherwise if a data point has more than one type of labels it is called multi-task learning

Multi-label classification

For example, a multi-label classification could be:

- 1 or 0 if a car is present or not in the image
- 1 or 0 if a person is present or not in the image
- 1 or 0 if a tree is present or not in the image
- 1 or 0 if a cat is present or not in the image

Number of features

It is better tu use only most relevant features but not fewer and missing relevant feature is bad for accuracy. Using irrelevant features instead wastes computation and might result in overfitting.

Model

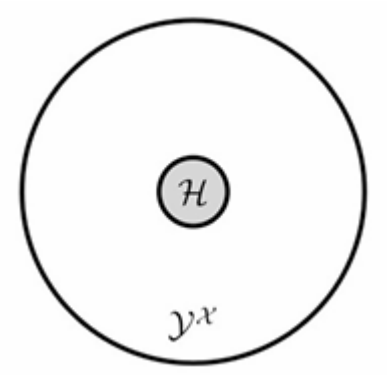
Machine learning

Machine learning means learning to predict the label y of a data point from its feature x, or learning an hypothesis $h \in H$ such that $h(x) = y$.

Hypothesis space

$h : X \to y \qquad h \in H$

they hypothesis space H is a /typically very small) subset of the (typically very large) set Y^X of all possible maps from feature space X into the label space Y.



Model to choose

The ide is to choose a large model in order to have flexibility to understand our date and patterns. It has to be small as well, in order to avoid overfitting.

Loss

Loss function

A loss function is a quantitative measure of prediction error obtained when using hypothesis h to predict label y′ of a data point with features x′.

To choose the right loss function you have to consider:

- statistical aspects
- computational aspects
- interpretation

Choosing a suitable loss function is often non-trivial.