

# Modelling data

Carl Edward Rasmussen

June 30th, 2016

# Key concepts

Non-technical introduction to terminology in modelling in machine learning:

- purpose of models
- central concepts
  - data
  - parameters
  - latent variables
- prediction and probabilistic predictions
- role of assumptions
- training and testing

# Purpose of models

Mathematical models are useful for many purposes, including

- making predictions. For example, in a time series model, we may want to predict the future from the past. Often predictions are inherently uncertain. In *probabilistic models* probabilities express the confidence of predictions
- generalize from observations in the training set to new test cases (interpolation and extrapolation)
- understanding and interpreting statistical relationships in the data
- evaluating the relative probability of hypothesis about the data
- compressing or summarising data
- generating more data, from a similar distribution as the training set

Different tasks require different models. Useful models focus on some aspects and neglect others, to *trade off accuracy with simplicity and interpretability*.

# Origin of a model

Models may originate from different sources, such as

- **first principles** For example, Newtonian mechanics is a model of planetary motion with a high degree of accuracy
- **observations, data** For example the annual production of timber per hectare of forrest, and its dependency on geographical and climatic factors may be modelled based on *data*.

Most practical models lie somewhere within the above spectrum, involving both first principles and data.

*Machine learning* is a broad term which covers the theory and practise of mathematical models which to a significant degree rely on data.

# Knowledge, assumptions and simplifying assumptions

Every model relies on (explicit or implicit) assumptions, such as

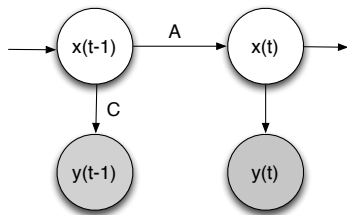
- **knowledge** For example, we know that a variable which measures the distance between two points must be non-negative.
- **assumptions** It could be assumed that income is independent of gender given age and profession. This assumption may be either true or false.
- **simplifying assumptions** We might assume that the response to a drug depends linearly on the dosage within some range. We don't believe that such an assumption is necessarily literally true, but rather that it is *good enough*, not to distort the uses we have of the model too much.

Probabilistic models may use *priors* to express knowledge (or beliefs) about aspects of the model.

Simplifying assumptions often facilitate use of the model. But, the conclusions drawn from a model are conditional on the assumptions being valid.

Practical modelling is therefore always a trade off between model expressivity and computational simplicity.

# Observations, parameters and latent variables



This time series model is imagined repeated for  $t = 1, \dots, T$ . It has *observations*  $y$  (shaded) for each time point and also *unobserved* or *hidden* or *latent variables*  $x$  and two sets of *parameters*  $A$  (for transitions) and  $C$  (for emissions).

To use the model we must decide what to do with all unobserved quantities. This is broadly known as *learning* or *training* a model. Options include *inference*, *estimation*, *sampling* and *marginalisation*.

Note, that the difference between *latent variables* and *parameters* is that the number latent variables grow with the number of observations (in this case, one for each time point), whereas the number of parameters remains constant.

# Practical modelling

The specification of a model includes the complete structure as well as all assumptions (and priors) used as well as any pre-specified parameters.

In practise, we need to be able to do the following tasks

- treat the unobserved quantities (training), including
  - the latent variables
  - the parameters
  - possibly some aspects of the structure of the model
- make predictions on test cases
- interpret the trained model, what insights is the model providing?
- evaluate the accuracy of model
  - note: accuracy on the training and test sets may differ systematically
- do model selection and model criticism: chose between different models, or between different variants of a model, what are limitations of the model?

All these tasks need to be solved either exactly or approximately, on a given budget of computation and memory.

# A common misunderstanding

The role of a model is to make predictions and provide insight into certain aspects of the data.

The role of a model is *not* to be a complete description of all aspects of the data (only the data itself does this).

From this perspective, it is clear that terms such as *true model* or *correct model* are *meaningless* in the context of machine learning.

*Essentially, all models are wrong, but some are useful.*

— George E. T. Box