Modelling data

Carl Edward Rasmussen

June 29th, 2016

Origin of a model

Mathematical models are useful for a number of different purposes.
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.

The purpose of a model

Mathematical models are useful for a number of different purposes.
These include, but are not restricted to

- making predictions. For example, in a time series model, we may want to predict the future from the past and the present
- generalize from the points in the training set to new test cases (interpolation and extrapolation).
- understanding and interpreting statistical relationships in the data
- evaluating the relative probability of various hypothesis about the data
- compressing or summarising data.

Because there are many uses for models, the usefulness of a model depends on which task we are trying to solve, different models are useful for different purposes.

Assumptions and simplifying assumptions

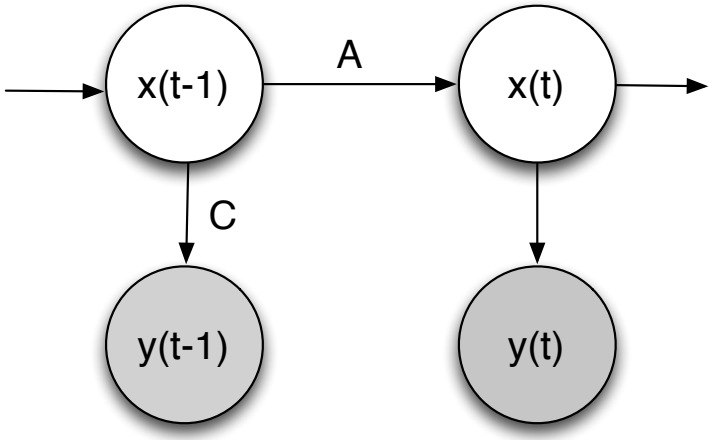
Every model relies on (explicit or implicit) assumptions.
There are two types of assumptions

insights For example, it might be known that a variable which measures the distance between two points must be non-negative.

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.

Often, simplifying assumptions give rise to huge mathematical simplifications.
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 on one hand and simplicity on the other.

Obervations, parameters and latent variables



This time series model is imagined repeated for $t = 1, \ldots, T$. 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. Possible options include inference, learning, estimation 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

What do we need to do: handle latent variables, handle parameters, model selection

some more about what is the difference between the data and a model. Avoid the term true model.