UvA-DARE (Digital Academic Repository)

Rules and associations : hidden Markov models and neural networks in the psychology of learning

Visser, I.

Publication date
2002

[Link to publication](#)

Citation for published version (APA):
Visser, I. (2002). *Rules and associations : hidden Markov models and neural networks in the psychology of learning*. [Thesis, fully internal, Universiteit van Amsterdam].

A Hidden Markov interpretations of neural networks

Abstract

Mike Pages manifesto makes a case for localist representations in neural networks, one of the advantages being ease of interpretation. However, even localist networks can be hard to interpret, especially when at some hidden layer of the network distributed representations are employed as is often the case. Hidden Markov models can be used to provide useful interpretable representations.

In his manifesto for the use of localist neural network models Mike Page mentions many advantages of such a scheme. One advantage is the ease of interpretation of the workings of such a network in psychologically relevant terms (section 7.6, Problems of Interpretation).

As Page justly remarks, a localist model does not imply that distributed representations are not used in any part of the model; rather a localist model is characterized by employing localist respresentations at some (crucial) points such as the output level of the network. More specifically he states that any entity that is locally represented at layer n of the hierarchy is sure to be represented in a distributed fashion at layer n-1 (section 2.6, So what is a localist model?). Why should the problem of interpretation not apply to these distributed representations at lower levels as well? I think it does, and its best to illustrate this with an example.

Following the work of Elman (1990), Cleeremans and McClelland (1991) used a simple recurrent network, SRN, to model implicit learning behavior using localist representations at both input and output layers, but a distributed representation at the hidden layer of the network. As they show in their paper the SRN model captures the main features of subjects performance by growing increasingly sensitive to the temporal context [of the current stimulus]. This sensitivity to the temporal context of stimuli is somehow captured by representations formed at the hidden layer of the network. In exactly what sense differences in temporal context affect activity at the hidden layer is unclear: what does a certain pattern of activity of the hidden layer units mean?

Visser et al. (2002) used hidden Markov models to characterize learning. By analyzing a series of responses it is possible to extract a hidden Markov model that is, in its general form, closely related to the sequence of stimuli that were used in the sequence learning experiment. In fact a hidden Markov model is a stochastic version of a finite state automaton, the kind of automaton used by Cleeremans

and McClelland (1991) to generate the stimuli for their implicit sequence learning experiment.

Such a procedure can also used in analyses of a neural network by having the network generate a series of responses or predictions. Using a version of the EM algorithm a hidden Markov model can be extracted from the network (see e.g. Rabiner, 1989). Extraction of a hidden Markov model of the network partitions the state space of the hidden layer of the network in discrete (hidden) states. This model is then interpretable in the sense that the presence or absence of connections between states indicates which sequences of stimuli are admissable and which are not; that is, the states can be regarded as statistically derived proxies of local representations. In addition, the extraction procedure does not rely on inspection of the activities at the hidden layer of the network as is done for example by Cleeremans et al. (1989) and Giles et al. (1992).

Extraction of hidden Markov models, and by implication of finite state machines, can in principle be applied to any neural network but is especially suitable for those that are used for modelling sequential behaviors. This is not to say that those networks should be replaced with HMMs. For example the work of Cleeremans and McClelland (1991) shows that their SRN model is very succesful in describing subjects behavior in implicit sequence learning. Although I strongly support Mike Pages manifesto for localist modelling it does not solve all problems of interpretation that arise in neural networks. Hidden Markov models are a highly useful tool to gain a better understanding of the internal workings of such networks in terms of proxies of local representations.