Conditioning

notes on "Explaining away in Weight Space" by Dayan and Kakade

Geoff Gordon
ggordon@cs.cmu.edu

February 5, 2001

Overview

HUGE literature of experiments on conditioning in animals

HUGE literature on optimal statistical inference

but relatively little overlap between them

which is a pity since conditioning is probably an attempt to approximate optimal statistical inference

Will describe research that attempts to make a connection

Conditioning

Most famous example: Pavlov's dogs

Learned to associate stimulus (bell) with reward (food)

Can get much more elaborate:

Name	Stimulus 1	Stimulus 2	Test
classical	B $\to$ R	—	B $\to \bullet$
sharing	B, L $\to$ R	—	B $\to \circ$, L $\to \circ$
forward blocking	B $\to$ R	B, L $\to$ R	B $\to \bullet$, L $\to \cdot$
backward blocking	B, L $\to$ R	B $\to$ R	B $\to \bullet$, L $\to \cdot$

$\bullet$ = expectation of reward $\quad \circ$ = weak expectation $\quad \cdot$ = no expectation

Statistical explanations

Simple models can explain some conditioning results

We'll discuss 2: gradient descent, Kalman filter

Models ignore (important) details:

- animals learn in continuous time
- animals have to sense stimuli and rewards
- animals filter out lots of irrelevant percepts
- ...

But they're still interesting as a simplification or an explanation of a piece of a larger system

Assumptions in both models

Trials presented as (stimulus, reward) pairs

Goal is to predict reward from stimulus

Learning is updating prediction rule

Stimulus $\in \mathbb{R}^n$ (in our case, 2 binary vars B and L)

Reward $\in \mathbb{R}$

Reward is linear fn of stimulus, plus Gaussian error

Gradient descent

Define

x_t input on trial t

y_t reward on trial t

w_t internal state (weights) after trial t

η arbitrary learning rate

Write expected reward $\widehat{y}_t = x_t \cdot w_t$, error $\epsilon_t = y_t - \widehat{y}_t$

Gradient descent model says:

$$w_{t+1} = w_t + \eta x_t \epsilon_t$$

Conditioning explained by gradient descent

In classical conditioning or sharing, +ve correlation between inputs and outputs causes relevant components of xy to be +ve, so those components of w become +ve

In forward blocking, stimulus 2 is explained perfectly by weights learned from stimulus 1, so no learning happens in phase 2 (error signal ϵ is 0)

Backward blocking

Gradient descent fails to explain backward blocking!

In stimulus 2 of backward blocking, the element of x_t corresponding to the light is always 0

So gradient descent predicts that the learned weight for the light won't change

Contradicted by experiments

Kalman filter explanation

Sutton (1992) proposed that classical conditioning could be explained as optimal Bayesian inference in a simple statistical model

The model:

- trial stimuli represented by vectors as before
- reward is linear function of stimuli plus Gaussian error
- in absence of information, weights of linear function drift over time in a Gaussian random walk

Inference in this model is called Kalman filtering

Kalman filter

Recall

x_t input on trial t

y_t reward on trial t

w_t weights after trial t

Assume

- $w_0 \sim N(0, \Sigma_0)$
- $w_{t+1} | w_t \sim N(w_t, \sigma^2 I)$
- $y_t \sim N(x_t \cdot w_t, \tau^2)$

Kalman filter cont'd

Write expected reward $\widehat{y}_t = x_t \cdot w_t$, error $\epsilon_t = y_t - \widehat{y}_t$

Calculate "learning rate" $\eta_t = 1/(\tau^2 + x_t^\top \Sigma_t x_t)$

Equations for new weights w_{t+1} and their covariance Σ_{t+1}:

$$\begin{aligned} z_t &= \Sigma_t x_t \\ w_{t+1} &= w_t + \eta_t \epsilon_t z_t \\ \Sigma_{t+1} &= \Sigma_t + \sigma^2 I - \eta_t z_t z_t^\top \end{aligned}$$

Comparison to GD

Update $w_{t+1} = w_t + \eta_t \epsilon_t z_t$ looks like GD, except:

η_t is a variable learning rate determined by variances of y_t and w_t

z_t instead of x_t plays role of input vector

Whitening

How to interpret z? (Recall $z = \Sigma x$)

z is a whitened or decorrelated version of x

To see why: fixed point of update for Σ is

$$\sigma^2 I = \eta z z^\top$$

which can only be true on average if z has spherical covariance

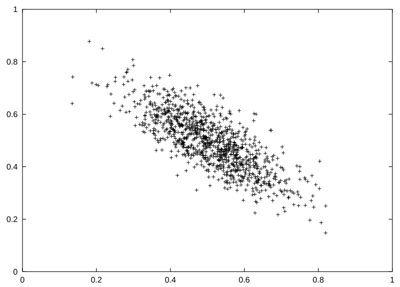
Conditioning

[Dayan&Kakade, 2000]: Kalman filter model explains all conditioning results from above

Classical, sharing, and forward blocking all work exactly as they did with the gradient descent model

But now backward blocking works too

Backward blocking



In sharing, $+$ve correlation between components of x_t makes off-diagonal elements of Σ become -ve in order to whiten

Interpretation: don't know whether it's B or L that's causing R

I.e., if we find out one weight is large, other must be small

I.e., evidence for B $\rightarrow$ R is evidence against L $\rightarrow$ R

"Explaining away"

Incremental version

D&K propose a network architecture using only fast computations which approximates the Kalman filter

Uses a whitening network from [Goodall, 1960] to get Σ and z, then z and error signal to get changes to w

Requires distribution of x_t to change only slowly (so whitening network converges)

Gets direction but not magnitude of update

Experimental results

D&K implemented the Kalman filter as well as the incremental network

Presented backward blocking stimulus: 20 trials of B,L $\to$ R, then 20 trials of B $\to$ R

Exact and incremental results qualitatively similar

Both show strong blocking effect

Discussion

What is essential difference between GD, KF?

- GD could simulate backwards blocking by using weight decay to “forget” $L \to \circ$
- But KF allows blocking and forgetting to happen on 2 different time scales (blocking is much faster)
- Works because KF can represent uncertainty separately for different directions in weight space

Discussion

What's important about KF?

- Gaussian assumption is clearly false, so that's not it
- Instead, idea that animals believe concept to be learned is changing over time

Improvements to KF:

- Use non-Gaussian distributions
- Use "punctuated equilibrium" rather than steady drift: concept is likely to stay same for a while, then change quickly to a new concept
- Use mixture models to remember previous concepts, switch between them

Conclusions

Simple statistical models can help explain experimental results on conditioning in animals (even if they gloss over important details)

Kalman filter is a better model than gradient descent: it constructs decorrelated features, so it can do backward blocking

Kalman filter is not best possible model, but provides guide to what characteristics a model needs to have