TD and Markov Chain Modelling for Sequence Prediction Tasks

COMP 767 Presentation

Charles C Onu

260663256

7 April, 2017

Objective

- To review the relationship between TD and Markov models in the setting of (supervised) outcome prediction from sequence/time-series data
- To investigate performance of TD(0) compared to direct generative modelling of a truly Markov chain

Supervised Learning (prediction from sequence data)

- Time-series/sequence-based prediction; e.g., chess positions and outcome; heart-rate readings and clinical outcome

$$x_1^1, x_2^1, ..., x_m^1, y^1$$

$$x_1^2, x_2^2, ..., x_m^2, y^2$$

$$...$$

$$x_1^n, x_2^n, ..., x_m^n, y^n$$

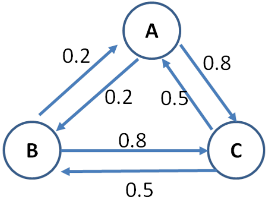
where $x_i^j \in R^k$

Relation to Model-free RL

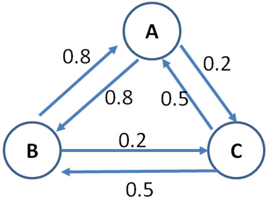
Sequence prediction is similar to the model-free reinforcement learning setting

- Similarity
  - Dynamics of underlying system is not known
  - We have actual example trajectories $x_1, x_2, ..., x_m$, and the final outcome y
- Difference
  - In RL: trying to figure how best to act in the system to accumulate 'good' outcomes (active learning)
  - In Supervised Learning: Given a trajectory already taken, what is the final outcome/reward - win or lose? rain or sun, survive or die

Experiment: an ergodic Markov chain



positive



negative

- Generates sequences of 5 time steps.
- Positive and negative examples are generated by different transition probabilities

Experiment: an ergodic Markov chain

Table 1: Transition probabilities for positive examples

	A	B	C
A	0	0.2	0.8
B	0.2	0	0.8
C	0.5	0.5	0

Table 2: Transition probabilities for negative examples

	A	B	C
A	0	0.8	0.2
B	0.8	0	0.2
C	0.5	0.5	0

Objective

- Given a set of sequences and outcomes generated by the system:

$$A, B, C, B, C, \mathbf{1}$$

$$C, A, B, A, B, \mathbf{0}$$

$$\ldots$$

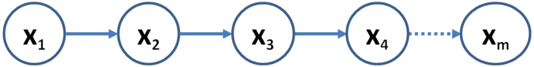
$$B, A, C, B, A, \mathbf{0}$$

- Predict the outcome of an observed trajectory:

$$C, A, C, B, C, \mathbf{?}$$

- Application: e.g., predicting a medical outcome (survive or not) given observed variable over time.
- Methodology
  1. Markov chain modelling (model-based)
  2. TD learning (model-free)

Methodology 1: Explicit Markov Chain Modelling (1)



- Parameters θ of the network are:
  - Start state distribution, $P(x_1)$
  - Transition probabilities, $P(x_t|x_{t-1})$ where $x_t \in \{A, B, C\}$
- Likelihood of a sequence:

$$L(x) = P(x_1) \prod_{t=1}^{m} P(x_t|x_{t-1})$$

Methodology 1: Explicit Markov Chain Modelling (2)

Learning

- Obtain maximum likelihood estimates (MLE) of parameters conditioned on positive θ_+ and negative examples θ_- respectively.
  - $\theta_+ : P_+(x_1), \quad P_+(x_t|x_{t-1})$
  - $\theta_- : P_-(x_1), \quad P_-(x_t|x_{t-1})$

Prediction

- For a new sequence x, select the most *likely* class.

$$\arg\max_c L(x|\theta_c)$$

where

$$L(x|\theta_c) = P_c(x_1) \prod_{t=1}^{m} P_c(x_t|x_{t-1})$$

Methodology 1: Explicit Markov Chain Modelling (3)

Result

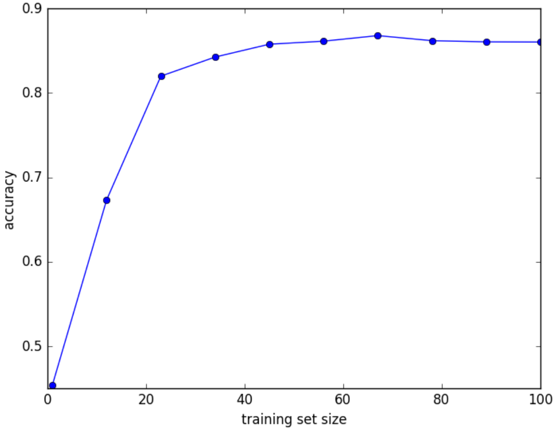


Figure 1: Accuracy vs training set size for Markov chain experiment

Methodology 2: TD Learning (1)

Parameter

- Values of each state V

Learning - TD(0)

- For t = 1 to m:

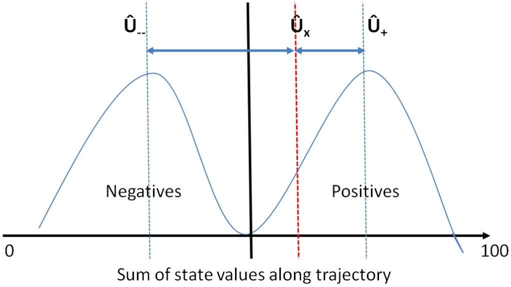
$$V(s_t) = V(s_t) + \alpha(r + \gamma V(s_t) - V(s_{t-1}))$$

- No discounting ($\gamma = 1$). At non-terminal states reward is 0. At terminal state reward is 0 or 1 depending on outcome.

Prediction

- For a new sequence C, A, C, B, C, we can obtain the sum of the state values: $U = 3V(C) + V(A) + V(B)$
- How do we predict the class from this real number? What is the threshold?

Methodology 2: TD Learning (2)



- Goal is to find the threshold that maximises the distance between the centres of the 2 distributions.
- Given a new sequence x, take the absolute distances:

$$d_+ = |U_x - \hat{U}_+|$$

$$d_- = |U_x - \hat{U}_-|$$

- Select the class c for which

$$\arg\min_c d_c$$

Methodology 2: TD Learning (3)

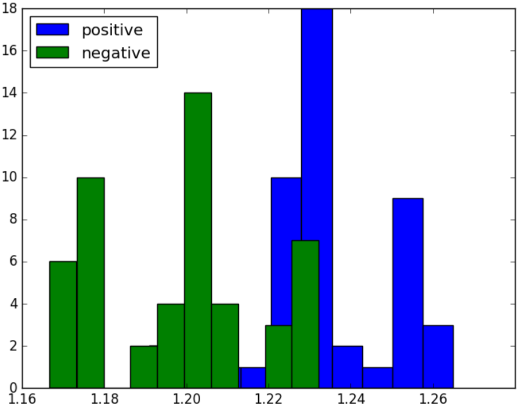


Figure 2: Histogram of state-value-sum of 100 training examples, computed after state values were learnt from same training data

Methodology 2: TD Learning (4)

Result

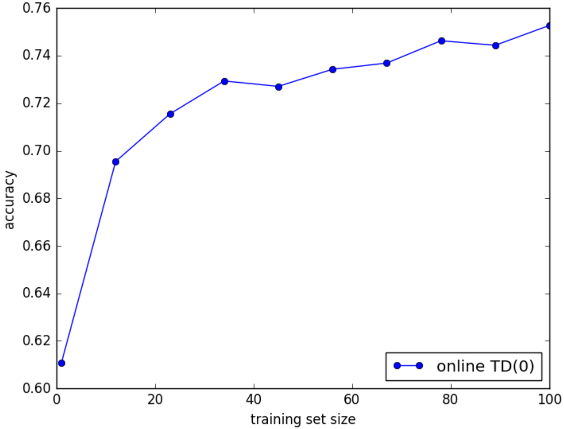


Figure 3: Accuracy vs training set size for online TD(0) experiment (averaged over 500 runs)

Online TD vs Markov chain modelling

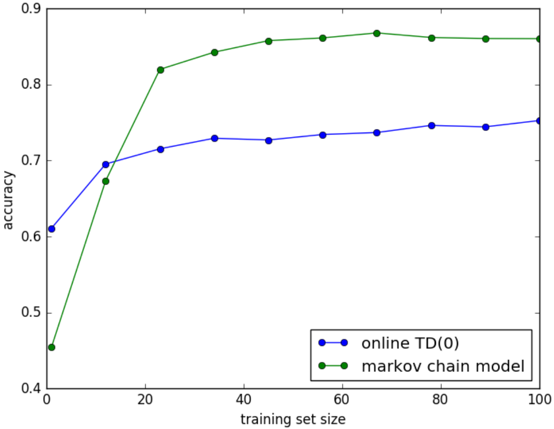


Figure 4: Accuracy vs training set size

But...

"batch TD(0) always finds the estimates that would be exactly correct for the maximum likelihood model of the Markov process" - Sutton and Barto, 2016

Online TD vs Batch TD vs Markov chain modelling

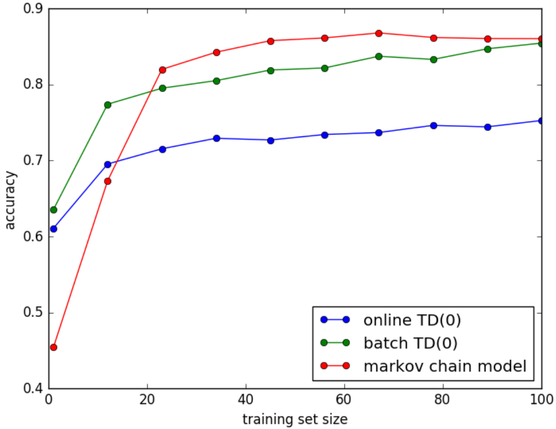


Figure 5: Accuracy vs training set size

Summary

- Goal was to determine how well TD would perform in a truly Markov system.
- We compare model-based (Markov chain modelling) vs model-free TD(0) learning.
- Batch TD(0) performed nearly as well as explicit Markov chain modelling. It may do just as well:
  - if we tune α (alpha was simply set to 0.01),
  - repeat presentation of each example more times (currently 1000 repeats)
  - learn better discriminator for the returns
- This suggests that indeed TD is approximating the Markov chain that defines the underlying system.
- TD (both online and batch) does better with little data than explicit Markov chain modelling. For more complex systems (e.g., larger state space) the training set size over which TD performs better could be larger.

References

- R. S. Sutton, "Learning to Predict by the Method of Temporal Differences" in Machine Learning 3, Boston:Kluwer Academic Publishers, 1988.
- E. Barnard, "Temporal-difference methods and Markov models" IEEE Transactions on Systems, Man, and Cybernetics, 23(2), 357–365, 1993.