Interpretability of Surrogate Models Produced by Viper

Otto Kaaij
Supervisor: Anna Lukina

TU Delft

Background

- There is a need for interpretable AI
- One possible approach:

Use imitation learning to extract interpretable surrogate models from black box oracles

Viper

Viper[1] is an imitation learning algorithm that prioritizes accuracy on critical states: states where making the wrong choice has the highest cost

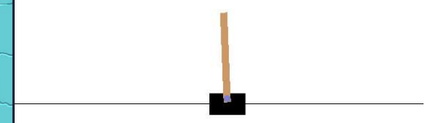
Goals

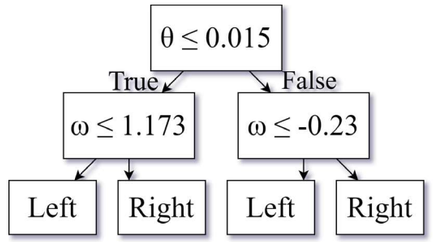
- Evaluate decision trees created by Viper on performance and interpretability
- Compare Viper to other imitation learning algorithms (Behavioral cloning (BC), Gail[2] and AggreVate[3])

Results

Cartpole

Keep the pole upright





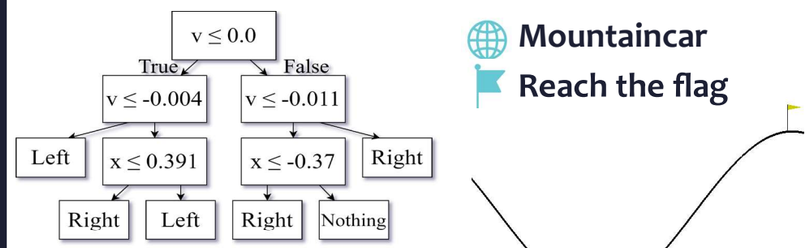
π* reward: 500±0.0 | π reward: 500±0.0

Interpretation

Move left if the pole is angled left, unless pole is rotating left quickly. Policy is symmetric to the right.

Mountaincar

Reach the flag



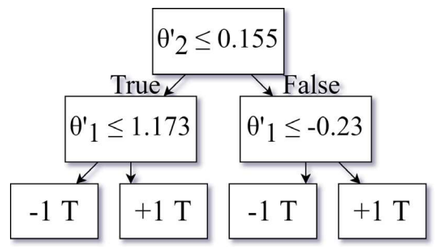
Interpretation

Accelerate in current direction to build momentum. Reverse just before standing still.

π* reward: -112.1±1.8 | π reward: -111.2±2.7

Acrobot

Swing above the line



π* reward: -85.9±17.1 | π reward: -84.3±20.81

Interpretation

'Kick' lower link to gain momentum. Build momentum until lower link flops over the line.

Algorithm Comparisons

- Viper, Gail and AggreVaTe all outperform baseline behavioral cloning
- Viper is the only algorithm that matches or improves on oracle performance on all three environments
- All algorithms produce similarly interpretable results, though Viper tends to produce smaller trees

Conclusion

- Viper produces high performance, interpretable decision trees for simple environments
- Performance is dependent on oracle quality, which cannot be expressed only in oracle reward
- We can use this interpretability to understand the policies and, in some cases, improve them

Future Work

- Evaluate more complex environments
- Evaluate other types of surrogate models
- Improve Viper by training decision trees more effectively and by cross-validating on interpretability
- Unify oracles for better comparisons

References

[1] Osbert Bastani, Yewen Pu, and ArmandoSolar-Lezama. "Verifiable reinforcement learning via policy extraction". In:arXiv preprintarXiv:1805.08328(2018)
[2] Jonathan Ho and Stefano Ermon. "Generative Adversarial Imitation Learning". In: Advances in Neural Information Processing Systems. Ed. by D.vLee et al. Vol. 29. Curran Associates, Inc., 2016.
[3] Stephane Ross and J Andrew Bagnell. "Reinforcement and imitation learning via interactive no-regret learning". In: arXiv preprint arXiv:1406.5979 (2014).
- Icons under CC by Milinda Courey, DinosoftLab and M. Oki Orlando from NounProject.com