A Supervised and Unsupervised Learning Extension to the Taxi RL Problem

Irfan Budi Satria[1] and Ikbal[2]

[1]Student ID: 2406514463, Faculty of Computer Science, Universitas Indonesia
[2]Student ID: 2406472826, Faculty of Computer Science, Universitas Indonesia

Abstract

This project aims to reproduce the classic Taxi reinforcement learning environment by Dietterich (1999), a taxi driver simulation aiming to train an agent to efficiently transport passengers between several locations. However, aside from the traditional RL objectives, we wish to extend the project by introducing two machine learning tasks:

- (1) a supervised learning model to predict cumulative earnings from optimal policies,
- (2) an unsupervised learning model for clustering the environment's state-space into meaningful regions.

1 Introduction

Reinforcement learning (RL) provides a framework for agents to learn optimal behaviors by interacting with an environment. Problems like the Taxi-v3 environment serve as simplified models for real-world tasks like ride-hailing and logistics. While RL focuses on decision-making through rewards, supervised learning excels at prediction, and unsupervised learning uncovers patterns without labels. This project aims to combine the three forms of ML, demonstrating how a learned policy can generate structured data for other machine learning tasks.

2 Related Work

The Taxi-v3 problem, introduced as part of OpenAI Gym's environments, is commonly used for teaching RL fundamentals. It represents a finite Markov Decision Process (MDP) with discrete states and actions.

There exists various examples of combining RL with supervised and unsupervised learning: supervised models often predict outcomes from state-action pairs, while clustering techniques have been used for environment abstraction, reward shaping, and exploration. Fei Feng (2020). Previous works include model-based RL with predictive modeling and hierarchical reinforcement learning through clustering.

3 Methodology

3.1 Environment Description

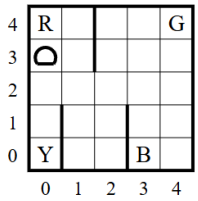

Figure 1: The Taxi Domain, taken from Dietterich (1999)

- Environment: Taxi-v3 (discrete grid-world)
- **State space**: 500 discrete states (taxi position, passenger location, destination). There are 25 taxi positions, 5 possible locations of the passenger (R, G, B, Y, and In Taxi), and 4 destination locations (R, G, B, Y).
- **Action space**: 6 discrete actions (South, North, East, West, Pickup, Drop-off)
- **Reward function**:
  - -1 per timestep,
  - $+20$ for successful dropoff,
  - -10 for illegal pickup/dropoff.

3.2 Data Requirements

- No external data needed for RL training (environment is simulated).
- Data for supervised/unsupervised learning will be generated:
  - **Supervised**: (state, action) $\rightarrow$ cumulative reward.
  - **Unsupervised**: environment states (possibly embedded).

3.3 Markov Decision Problem Formulation

- $\mathcal{S}$: finite set of states (500 states).
- $\mathcal{A}$: finite set of actions (6 actions).
- $P(s'|s,a)$: transition dynamics provided by environment.
- $R(s,a)$: reward function as described.
- γ: discount factor (e.g., 0.95).

3.4 RL Training Setup

- **Algorithm**: Q-learning based on Dietterich (1999)
- **Hyperparameters**:
  - Learning rate (α): 0.1
  - Discount factor (γ): 0.95
  - Exploration rate (ϵ): decaying from 1.0 to 0.01
- **Training episodes**: $\sim$10,000

3.5 Supervised Learning Setup

- **Task**: Predict cumulative earnings (reward) given (state, action).
- **Model**: Regression model (e.g., Random Forest).
- **Input**: Encoded state features + action.

3.6 Unsupervised Learning Setup

- **Task**: Cluster states into function areas (Dropoff/Pickup Zones).
- **Model**: K-Means clustering.
- **Input**: Encoded state features + action.

4 Initial Results and Discussion

Initial experiments will focus on replicating standard Q-learning performance on the Taxi-v3 environment. Once an optimal policy is learned, datasets for supervised and unsupervised tasks will be generated. Preliminary expectations:

Regression models will predict reward with reasonable error after feature engineering.

Clustering will reveal logical groupings (e.g., passenger pickup areas vs random areas).

5 Conclusion

This project explores how reinforcement learning can be expanded with supervised and unsupervised learning tasks, making the RL problem a data generator for further ML Applications. The approach of further analysis with Supervised and Unsupervised Learning will surely enhance our understanding of the taxi environment. Future work could include testing transfer learning between environments and dynamic reward adjustment based on clustering.

References

Dietterich, T. G. (1999). Hierarchical reinforcement learning with the maxq value function decomposition. Technical report, Department of Computer Science, Oregon State University,, Corvallis, OR 97331. Available online at http://arxiv.org/abs/cs/9905014v1. (pages 1 and 2)

Fei Feng, Ruosong Wang, W. Y. S. S. D. L. F. Y. (2020). Provably efficient exploration for reinforcement learning using unsupervised learning. Technical report, University of California, Los Angeles. Available online at https://arxiv.org/abs/2003.06898. (page 1)