Machine learning excersice

Alexander van Heteren

I used Claude 4.5 sonnet to look at the experiments and give insights on the outcomes.

Introduction

This project implements Q-learning on the FrozenLake environment from Gymnasium to study how different hyperparameters affect agent learning. The goal was to understand the exploration-exploitation tradeoff and test various reward shaping strategies.

Methodology

I implemented a tabular Q-learning agent that learns optimal policies through trial and error. The agent updates Q-values using: $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max Q(s',a') - Q(s,a)]$

The experiments were run on FrozenLake 4x4 environment for 10,000 episodes each, with evaluation every 100 episodes.

Experiments & Results

Experiment 1: Learning Rate Comparison

Tested three learning rates (α = 0.1, 0.3, 0.7). Results showed that lower learning rates (0.1-0.3) achieved ~70% success rate, while high learning rate (0.7) completely failed to learn (0% success). This demonstrates that too high learning rates cause unstable Q-value updates that prevent convergence.

Experiment 2: Stochastic vs Deterministic Environment

Compared slippery (stochastic) and non-slippery (deterministic) versions. The deterministic version learned faster and reached higher success rates because actions always had predictable outcomes. The stochastic version showed more variability in learning curves due to random state transitions.

Experiment 3: Reward Shaping

Tested different penalty configurations. Adding small step penalties (-0.01) encouraged the agent to find shorter paths. Large hole penalties (-1.0) made the agent more cautious but didn't significantly improve final performance compared to no shaping.

Experiment 4: Exploration Strategies

Compared epsilon-greedy with Boltzmann (softmax) exploration. Epsilon-greedy showed more stable learning curves. Boltzmann with high temperature explored more initially but sometimes got stuck in suboptimal policies. Both strategies eventually converged to similar performance.

Experiment 5: Discount Factor

Tested γ values of 0.9, 0.95, and 0.99. Higher discount factors (0.99) performed better because they valued future rewards more, which is important in FrozenLake where the goal is several steps away.

Q-Table Analysis

The Q-table heatmaps show clear "craving" behavior - states near the goal have significantly higher Q-values. The optimal policy arrows show the agent learned to navigate around holes toward the goal. States adjacent to holes have lower values, showing the agent learned to avoid them.

Conclusions

- Learning rate is critical
- FrozenLake makes learning slower and more variable
- Reward shaping speeds up learning but needs tuning
- Epsilon-greedy is more reliable than Boltzmann for this environment

Code https://github.com/einstein43/RL