Human-centered Artificial Intelligence

2024/2025

Worksheet #6:
Human in the loop machine learning (part II):

Luís Macedo

6.1 Topics

- Learning how to behave from human demonstrations, feedback, and natural language
  - Learning from demonstration – AI agents learning to do what humans want from human demonstrations
    - Direct (behaviour cloning/imitation learning)
    - Indirect (preference learning: learning the reward function - Inverse reinforcement learning)
  - Reinforcement learning from human feedback (comparison-based feedback, reward modeling, corrective feedback)
  - AI agents learning to do what humans want from natural language
- Predicting/Inferencing human beliefs, desires, intentions, goals, and actions; cognitive modeling, personalisation

6.2 Pre-class Readings

Look at the following resources:

- https://livebook.manning.com/book/human-in-the-loop-machine-learning/chapter-1/
- https://www.youtube.com/watch?v=VMp6pq6_QjI
- https://www.youtube.com/watch?v=GdTBqBnqhaQ
- https://www.youtube.com/watch?v=Lu56xVlZ40M
- https://www.youtube.com/watch?v=NOLAwD4ZTW0 (*Robot learning with a joystick using RL+help from humans (Direct behaviour cloning)*)
- https://www.youtube.com/watch?v=ta9SkcJdSBA (*Robot Learning from Demonstration by Averaging Trajectories - Making Coffee*)
- http://www.scholarpedia.org/article/Robot_learning_by_demonstration (*Robot learning by demonstration*)
- https://www.youtube.com/watch?v=tcFLwgqO7G0 (*Stanford Seminar - The Next Generation of Robot Learning*)

- https://www.youtube.com/watch?v=uC-HoL6N7rI (*Robot Learning from Demonstration: From Mimicking to Emulation*)
- https://www.youtube.com/watch?v=CW1s6psByxk (*Talk: Sergey Levine, UC Berkeley - Learning Dynamic Manipulation Skills*)

6.3 Theoretic-Practical Exercises

6.3.1 Overview of HitL Concepts

Question 6.1 What is **Human-in-the-Loop Machine Learning**, and how does it differ from traditional machine learning?

Question 6.2 Why is human feedback valuable in training AI models for complex, real-world tasks?

Question 6.3 What are the main benefits of incorporating human feedback in AI learning systems?

Question 6.4 What are some potential risks or limitations of Human-in-the-Loop learning systems?

Question 6.5 How does Human-in-the-Loop learning help in aligning AI systems with human values and ethical considerations?

6.3.2 Learning from Demonstration (LfD)

Overview of Learning from Demonstration

Question 6.6 Define **Learning from Demonstration (LfD)** and explain its significance in teaching AI models.

Question 6.7 What are the differences between **direct** (behavior cloning) and **indirect** (inverse reinforcement learning) methods in Learning from Demonstration?

Question 6.8 How does **behavior cloning** work, and what are some common challenges associated with it?

Question 6.9 Explain **Inverse Reinforcement Learning (IRL)** and how it enables AI to infer the objectives behind human actions.

Question 6.10 Give an example of how Learning from Demonstration could be used in robotics. What challenges might arise?

6.3.3 Reinforcement Learning from Human Feedback (RLHF)

RLHF Overview and Techniques

Question 6.11 What is **Reinforcement Learning from Human Feedback (RLHF)**, and how does it enhance traditional reinforcement learning?

Question 6.12 Describe the three types of human feedback commonly used in RLHF: Comparison-Based Feedback, Reward Modeling, and Corrective Feedback.

Question 6.13 How is RLHF used to train language models like ChatGPT? What is the role of human feedback in this context?

Question 6.14 Discuss some advantages and potential pitfalls of using RLHF in high-stakes applications like autonomous vehicles.

Question 6.15 In what scenarios is RLHF particularly useful? Why might RLHF be preferred over traditional reinforcement learning in these cases?

6.3.4 Learning from Natural Language

Instruction and Reward Shaping

Question 6.16 How can natural language be used as an instructional medium for training AI?

Question 6.17 Explain **language-based reward shaping** and give an example of its application.

Question 6.18 How does **instruction-based reinforcement learning** work, and in what contexts might it be useful?

Question 6.19 Discuss the challenges of using natural language for AI instruction. What are the potential limitations?

Question 6.20 How could natural language instructions be combined with RLHF to improve AI performance and alignment?

6.3.5 Predicting Human Beliefs, Desires, and Intentions

Cognitive Modeling and Personalization

Question 6.21 What is cognitive modeling in the context of Human-in-the-Loop learning?

Question 6.22 How can AI systems predict human beliefs, desires, and intentions to improve personalization?

Question 6.23 Describe **goal recognition models** and explain how they might be applied in real-world AI systems.

Question 6.24 What are some ethical considerations in designing AI systems that predict human beliefs or intentions?

Question 6.25 Provide examples of how personalized AI systems that anticipate user needs might enhance user experience in e-commerce or healthcare.

6.3.6 Applications and Case Studies

Real-World Applications of RLHF, LfD, and Cognitive Modeling

Question 6.26 Describe some real-world applications of Learning from Demonstration (LfD) and RLHF in robotics.

Question 6.27 How are RLHF and LfD used in autonomous driving systems to improve safety and decision-making?

Question 6.28 In what ways can RLHF be applied in gaming to improve AI performance and user experience?

Question 6.29 Explain how cognitive modeling could improve the functionality of virtual assistants or customer service bots.

Question 6.30 Discuss the role of human feedback in enhancing language models. How does this feedback contribute to their adaptability and ethical considerations?

6.3.7 Critical Analysis and Future Directions

Long-Term Implications and Research Directions

Question 6.31 What are the main challenges associated with scaling Human-in-the-Loop Machine Learning for complex AI systems?

Question 6.32 Discuss the implications of biased human feedback in HitL systems. How might this bias affect AI performance?

Question 6.33 What are the potential long-term impacts of relying on human feedback to guide AI development?

Question 6.34 What areas of research are currently active in enhancing HitL, LfD, and RLHF? What improvements might we expect in the future?

Question 6.35 How might future advances in Human-in-the-Loop learning affect fields like healthcare, autonomous systems, or creative industries?

6.3.8 Theoretical-practical exercises

Question 6.36 Consider now the situations in the next videos. They represent a different form of interaction/collaboration between an AI agent and humans. What are the differences to the previous videos? What information is given by humans to the robots?

http://www.scholarpedia.org/article/Robot_learning_by_demonstration

https://www.youtube.com/watch?v=ta9SkcJdSBA

Question 6.37 (*Adapted from the Normal Exam of AI-2020 edition*) For each of the applications/scenarios described below, indicate which technology or combination of technologies (Reinforcement Learning (RL), or Learning from Demonstration/Apprenticeship Learning – Behavior Cloning (IL-BC), or Learning from Demonstration/Apprenticeship Learning – Inverse RL (IL-IRL)) is best suited so that the best performance is achieved. Justify your answer.

- Consider an E-learning system in which an Artificial Intelligent Personal Assistant (AIPA) is to be integrated. This AIPA should build learning paths (sequence of learning activities – practical, theoretical exercises, classes, seminars, etc.) personalized for each student, i.e., it should obtain a function that specifies what the student should do in certain circumstances (student status).
- Consider a domestic robot. The main task is to cook.
- Consider the scenario of an intelligent robot that supports the elderly in a Nursing Home. The main task performed by the robot is the distribution of drugs to each user according to the prescribed doses, at the recommended times.
- Consider the scenario of an intelligent robot that supports the elderly in a Nursing Home. The main task is to entertain the elderly by telling them jokes and funny stories.

6.3.9 Reinforcement Learning

Question 6.38

How can we build an autonomous artificial agent that is able by himself to learn making decisions with no supervision and by interacting with the environment?

As an example, take a look at:
https://www.youtube.com/watch?v=NOLAwD4ZTW0 (robot)
https://www.youtube.com/watch?v=VMp6pq6_QjI (parking)

Question 6.39

Consider the grid world environment (slides 7–12). Can the sequence of actions

$$Up, Up, Right, Right, Right$$

take the agent in terminal state (3,4)? What's the probability? Can the sequence reach the goal in any other way?

Question 6.40

Consider a Markov Decision Process (MDP).

- List MDP main features.
- How is it formally defined?
- What's a policy and an optimal policy?
- What's the difference between Rewards and Utilities/Values?
- What's the main goal of a MDP?
- Analyse and understand the impact on the optimal policy of considering different reward values in the states of the environment. E.g.:
  - Why is the agent heading straight into (2,4) from it surrounding states when the reward is -1.6?
  - Why is the agent heading straight into the obstacle from (2,3), when $-0.0218 < reward < 0$?
  - Why the agent avoids the terminals when reward > 0 ?

Question 6.41 What is Value Iteration? What's the difference between (i) the expected value of following policy PI in a state s (V()), and (ii) the expected value of performing an action a in a state s, and then following policy PI (Q-value) ?

Question 6.42 (adapted from the Normal Exam of AI 2018/2019)

Consider the grid of slide 37 representing a MDP environment and assume that the current V*(s) values (the long run utility values) at step k of the Value Iteration algorithm, i.e., V(K)(s), are those represented in each cell.

Assume also that R(s) = -0.04, for s non-terminal, R(s) = -1 for cell (2,4), R(s) = +1 for cell (3,4), and the discount factor Y=1. Consider also that the agent moves in the above grid via actions Up, Down, Left, Right. Each action has 0.8 probability to reach its intended effect, 0.1 probability to move at right angles of the intended direction. If the agent bumps into a wall, it stays there.

- Compute the Q-value for the cell of row 3 and column 1: Q*((3,1),a), for a in {Up,Down,Left,Right}.
- Compute the optimal policy for the same cell, $n*((3,1))$, and update the value of $V*((3,1))$.

Question 6.43 Let's see now how all that began.

Consider now the initial conditions. Consider R(s) = -0.04, for s non-terminal, R(s) = -1 for cell (2,4), R(s) = +1 for cell (3,4), and the discount factor Y=1. Suppose that we start with values V(0)(s) that are all 0.

- Compute the Q-value for the cell of row 1 and column 1: $Q*((1,1),a)$, for a in Up,Down,Left,Right.
- Compute the optimal policy for the same cell, $n*((1,1))$, and update the value of $V*((1,1))$.

Question 6.44 Consider now the initial conditions. Consider R(s) = -0.04, for s non-terminal, R(s) = -1 for cell (2,4), R(s) = +1 for cell (3,4), and the discount factor Y=1. Suppose that we start with values V(0)(s) that are all 0.

- Compute the Q-value for the cell of row 1 and column 1: Q*((3,3),a), for a in {Up,Down,Left,Right}.
- Compute the optimal policy for the same cell, $n*((3,3))$, and update the value of $V*((3,3))$.

Question 6.45 Consider now the initial conditions. Consider R(s) = 0, for s non-terminal, R(s) = -1 for cell (2,4), R(s) = +1 for cell (3,4), and the discount factor Y=0.9. Suppose that we start with values V(0)(s) that are all 0.

- Compute the Q-value for the cell of row 1 and column 1: Q*((3,3),a), for a in {Up,Down,Left,Right}.
- Compute the optimal policy for the same cell, $n*((3,3))$, and update the value of $V*((3,3))$.

Bibliography