AI project themes

2021/2022

Luís Macedo

Notes: The following list of project themes involve to some extent all the three broad modules of this course of AI, namely Multi-agent systems, Uncertain Representation and Reasoning, and Learning. The goal is that the solution/implementation of the project relies largely on the contents taught in the lectures of this course plus some research on specific fields, in which you are expected to especialise,. Any doubts or inquires concerning each theme don't hesitate to contact me via Slack workspace or go to PLs classes.

Project 1

Collaborative Exploration of Unknown Environments with Teams of Agents

Exploration may be defined as the process of selecting and executing actions so that the maximal knowledge of the environment is acquired. The result is the acquisition of models of the physical environment. Unfortunately, exploring unknown environments requires resources from agents such as time and power. Thus, there is a trade-off between the amount of knowledge acquired and the cost to acquire it. The goal of an explorer is to get the maximum knowledge of the environment at the minimum cost (e.g.: minimum time and/or power). Therefore, strategies that minimize the required amount of these resources and maximize knowledge acquisition have been pursued. Several exploration strategies may be successfully applied. They might be regarding algorithms of selecting the best viewpoints, i.e., on how single agents select their way in the environment, or regarding multiple agents, i.e., the multiagent system architecture, their hierarchy, if they are working on teams, how those teams collaborate, cooperate and negotiate in the construction of the map of the environment. For this reason, this project might be split into several ones, if there are more than on student (or group) applying to it. This work is a prosecution of [Macedo, Tavares, Gaspar, Cardoso, EPIA2011] and the later improvemnts, focusing on the method for classifying the objects, based on Machine Learning algorithm such as Naïve Bayes. The software developed by [Macedo, Tavares, Gaspar, Cardoso, EPIA2011] (https://www.cisuc.uc.pt/publication/show/3411) will be the basis of this project. Furthermore, the following aspects should be addressed:

a) More kinds of objects with more features and class labels (perhaps a more attractive environment)

b) The experimental set up should be improved with more details about the object attributes used or the frequency distribution of the object classes

Various variants can be considered within this project such as those resulting considering different architectures for the multi-agent system (master-slave, flat, etc.), or integrating it with recent advances in Active Learning, Reinforcement Learning, Multiagent Reinforcement Learning and/or Imitation Learning, or Explainable AI (explaining the decisions made by the agents).

Toolkits: Mason, etc.

Project 2

Human in the Loop: Deciding what to learn - a new Active Learning Algorithm

The key idea behind active learning is that a machine learning algorithm can perform better with less training if it is allowed to choose the data from which it learns. An active learner may pose "queries,"usually in the form of unlabeled data instances to be labeled by an "oracle" that already understands the nature of the problem (e.g., a human annotator which is supposed to be an expert in the subject). This sort of approach is well-motivated in many modern machine learning and data mining applications, where unlabeled data may be abundant or easy to come by, but training labels are difficult, time-consuming, or expensive to obtain. There are already several algorithms, such as those belonging to pool-based active learning, in which the learner (i.e., the classifier) has access to a pool of unlabeled instances (e.g., emails) and can request labels (spam/not spam) for instances likely to be most helpful in building a predictive model. In uncertainty sampling, the classifier requests examples about whose class membership it least certain.

The goal of this project is to develop a new algorithm to be incorporated in artificial agents with the aim of decreasing the amount of information (input-output pairs) considered in the agents own decisions or added to its belief store, and/or decreasing the amount of information provided by the agents to humans. Such artificial agents can be used in several domains such as medical domain, spam detection (the application domain will be selected by the student). For instance, if integrated with a supervised machine learning classification system in the medical domain (e.g., when a medical doctor has to label the training examples of a machine learning system for disease categorization), such an agent would select a small set of examples in order to avoid the burden of labeling thousands of examples by a medical doctor ("I need to know all these labelled instances, but, since it is too much effort for you to give me all this information, please give me part of it because this part that I selected carefully is expected to cover the most of the subject").

Various variants can be considered within this project such as those resulting from considering different data sets, single versus multiple human experts, different information imetris (including intrinsic motivation models, e.g., based on surprise value of information – see [Macedo et al 2004, 2001, 2009]), Explainable AI (explaining the decisions made by the learning agents), etc..

Toolkits: JCLAL (https://github.com/ogreyesp/JCLAL), Libact (https://github.com/ntucllab/libact), etc.

Project 3

Learning Interesting New Classes - Active learning under the open set assumption

Traditional machine learning systems classify instances relying on a closed world assumption. The classes that the learning agent sees in training are what it will see in testing (no new objects or classes can appear in testing) (Fei and Liu 2016). However, in many real world problems, there is an open set of classes, which means the system should consider in some situations the possibility of labelling new unlabelled instances with new classes that are not in the set of those already known from the data set it holds. The issue for the artificial intelligent system is on detecting those situations that are of interest to be learned (because they might contribute with rich new information) and act accordingly, either by classifying by itself the new instance in the light of an open set of classes or by alerting and/or asking help to humans to do that. This problem is called open-world learning (or open-world classification). Apart from detecting the unseen classes, open-world learning should also incrementally or continually learn the new classes.

The goal is to design a learning algorithm that can classify data of the known/seen classes into their respective classes and also to detect instances from unknown/unseen classes. More precisely, the goal of this project is to solve this problem relying on selecting for labelling by an oracle those unlabelled instances that are expected to be more cognitively interesting because they represent situations in which the system is confused and/or because they might contribute with new, surprising information. A possible approach may involve evaluating the informativeness of an unlabelled instance in terms of the uncertainty, novelty and surprise of its possible class labels.

Toolkits: JCLAL (https://github.com/ogreyesp/JCLAL), Libact (https://github.com/ntucllab/libact), etc.

Project 4

Multi/Single-agent Reinforcement Learning

Reinforcement learning (RL) [Sutton,98] is a successfull Artificial Intelligence technique for learning to act in sequence. It relies on the mathematical formalism of Markov Processes, which means the definition of a problem requires defining the set of states, set of actions, the transition model, and reward model. The goal is to learn a policy, i.e., how the agent should act in specific states. It has been successfully applied in driverless cars, games, etc. Even though the success of RL in making agents learn to act in complex environments by trial and error, there are still ways of making this Machine Learning branch evolve.

The goal of this project is to replicate or reproduce an implementation of RL in a specific domain, and if possible to inovate so that the State of the Art is advanced. Connections with Multiagent Systems, Imitation Learning and Explainable AI can be established.

Toolkits: open source software https://awesomeopensource.com/projects/reinforcement-learning (see also https://awesomeopensource.com/project/ugurkanates/awesome-real-world-rl)

Project 5

Imitation Learning (Learning from Demonstrations) - Behaviour Cloning; Inverse Reinforcement Learning

Reinforcement Learning (RL) [Sutton 98] is a successfull Artificial Intelligence technique for learning to act in sequence. It relies on the mathematical formalism of Markov Processes, which means the definition of a problem requires defining the set of states, set of actions, the transition model, and reward model. The goal is to learn a policy, i.e., how the agent should act in specific states. It has been successfully applied in driver less cars, games, etc. However, to work well, they require lots of data, which is most of the times not available in several domains. Imitation Learning [Son19] is a type of machine learning closely related to RL. It is especially relevant when there is no reward function (Imitation Learning is often preferred over RL when such a reward function is difficult to specify (e.g., act friendly), or when the reward function is sparse and difficult to optimize directly) or policy available as it allows to learn either the reward (through the use of Inverse Reinforcement Learning [Ng 00], a sub type of Imitation Learning) or the policy from a demonstrator or trainer, which in this case is a role played by a teacher.

The goal of this project is to replicate or reproduce an implementation of Imitation Learning in a specific domain, and if possible to inovate so that the State of the Art is advanced. Various links to other theoretical concepts of the AI course can be found, including Inverse Reinforcement Learning, Explainable AI, and Ethics and Values.

Toolkits: open source from https://awesomeopensource.com/projects/imitation-learning (see also https://awesomeopensource.com/project/kristery/Awesome-Imitation-Learning; https://awesomeopensource.com/project/kristery/Awesome-Imitation-Learning#tutorials-and-talks; ´https://awesomeopensource.com/project/yrlu/irl-imitation)

Project 6

Learning Ethics in AI agents

Embedding ethics into AI systems remains an outstanding challenge. We all eventually want AI to behave morally, but so far we have no way of measuring a system?s grasp of general human values (Müller, 2020). The demand for ethical machine learning (White House, 2016; European Commission, 2019) has already led researchers to propose various ethical principles for narrow applications. To make algorithms more fair, researchers have proposed precise mathematical criteria. However, many of these fairness criteria have been shown to be mutually incompatible (Kleinberg et al., 2017), and these rigid formalizations are task-specific and have been criticized for being simplistic.

Through their work on fairness, safety, prosocial behavior, and utility, researchers have in fact developed proto-ethical methods that resemble small facets of broader theories in normative ethics. Fairness is a concept of justice, which is more broadly composed of concepts like impartiality and desert. Having systems abide by safety constraints is similar to deontological ethics, which determines right and wrong based on a collection of rules. Imitating prosocial behavior and demonstrations is an aspect of virtue ethics, which locates moral behavior in the imitation of virtuous agents. Improving utility by learning human preferences can be viewed as part of utilitarianism, which is a theory that advocates maximizing the aggregate well-being of all people. Consequently, many researchers who have tried encouraging some form of ?good? behavior in systems have actually been applying small pieces of broad and well-established theories in normative ethics.

The goal of this project is to develop a model for AI ethics.

Toolkits: open source from https://awesomeopensource.com/projects/imitation-learning (see also https://awesomeopensource.com/project/kristery/Awesome-Imitation-Learning; https://awesomeopensource.com/project/kristery/Awesome-Imitation-Learning#tutorials-and-talks; ´https://awesomeopensource.com/project/yrlu/irl-imitation)