Robot Learning

Exercise 1 – Imitation Learning

Release date: Wed, 13 September 2023 - **Deadline for Homework: Wed, 27 September 2023 - 23:59**

For this exercise you need to submit a **.zip** folder containing your report as a **.pdf** file (up to 3 pages), your pre-trained model as a **.t7** file and your code (namely the files `network.py, training.py, imitations.py`). Please use the provided code templates for these exercises. The given `main.py` file will be used for evaluating your code.
Changes to this main file, to the gym environment or additionally installed packages will not be considered. Comment your code clearly, use docstrings and self-explanatory variable names and structure your code well.

1.1 Network design (1+2+2+2+2+1 Points)

Let $\mathcal{S}$ be the state space and $\mathcal{A}$ the action space. The gym environment encodes a transition function $T : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ that maps a (state, action) - pair to a new state. It also provides a reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ which we will only use for evaluation this time.
The aim of this exercise is to design a network that learns a policy $\pi_\theta : \mathcal{S} \to \mathcal{A}$, parameterized by θ, to predict the best action for a given state. We formulate it as a supervised learning problem, where the objective is to follow the observed imitations of an "expert" driver.

a) Some imitations from an expert are given in `data/teacher`. Implement the function `load_imitations` in `imitations.py`. Given the folder of the imitations, it should load all `observation` and `action` files into two separate lists `observations` and `actions` which are returned.

b) The module `training.py` contains the training loop for the network. Read and understand its function `train`. Why is it necessary to divide the data into batches? What is an epoch? What do lines 43 to 48 do? Please answer shortly and precisely.

c) To start with, we formulate the problem as a classification network. Have a look at the actions provided by the expert imitations, there are three controls: steer, gas and brake. Which values do they take? Since they are not independent (accelerate and brake simultaneously does not make sense), it is reasonable to define classes of possible actions, like {steer_left}, {}, {steer_right and brake}, {gas} and so forth.
Define the set of action-classes you want to use and complete the class-methods `actions_to_classes` and `scores_to_action` in `network.py` The former maps every action to its class representation using a one-hot encoding and the latter retrieves an action from the scores predicted by the network. Lastly, implement the loss function `cross_entropy_loss` in `training.py` to calculate the training loss for a given pair of predicted and ground truth classes.

d) Design and implement an easy first network architecture in `network.py`. Start with 2 to 3 2D convolution layers on the images, followed by 2 or 3 fully connected layers (linear layers) to extract a 1D feature vector. Let each layer be followed by a ReLU as the non-linear activation (see code snippets below).
The output of the network should function as a controller for the car and predict one of the action-classes for each given state. At the end of the network, add a softmax layer to normalize the output. We can then interpret it as a probability distribution over the action-classes.

```
1          torch.nn.Sequential(
2             torch.nn.Conv2d(in_channels, out_channels, filter_size, stride=*arg),
3             torch.nn.LeakyReLU(negative_slope=0.2))
```

```
1          torch.nn.Sequential(
2             torch.nn.Linear(in_size, out_size),
3             torch.nn.LeakyReLU(negative_slope=0.2))
```

e) Implement the forward pass for your network, which is the function `forward` in `network.py`. Given an observation, it should return the probability distribution over the action-classes predicted by the network. You can decide whether you want to work with all 3 color channels or convert them to gray-scale beforehand. Motivate your choice briefly.
Train your network by running `python3 main.py train`. Afterwards, enjoy watching its performance by running `python3 main.py test` on your local machine. Can you achieve better results when changing the hyper-parameters? Can you explain this?

f) `imitations.py` provides some code to record new imitations. Complete the function `save_imitations` and start driving yourself by running `python3 main.py teach` on your local machine. What is 'good' training data? Is there any problem with only perfect imitations?

1.2 Network Improvements (2+2+2+2+2 Points)

Now that your network is up and running, it's time to increase its performance! Each of the following tasks adds to its architecture. It is up to you to choose which of them you use for participating in the competition. However, all subtasks need to be answered. Evaluate and compare different methods always on the same training data (no matter whether that is the provided or self-recorded data or a mix of both).

a) **Observations.** The training data of the network actually contains more information than just the image from the car in the environment. Look at the class method `extract_sensor_values` in `network.py`. What does it do? Incorporate it into your network architecture. How does the performance change?

b) **MultiClass prediction.** Design a second network architecture that encodes a multi-class approach by defining 4 binary classes that represent the 4 arrow keys on a keyboard and stand for: steer right, steer left, accelerate and brake. Since those don't all exclude each other, let the network learn to predict 0 or 1 for each class independently.
You will need to implement another loss function and might find a sigmoid-activation function useful. Again, compare the results to the previous classification approach.

c) **Classification vs. regression.** Formulate the current problem as a regression network. Which loss function is appropriate? What are the advantages / drawbacks compared to the classification networks? Is it reasonable to use a regression approach given our training data?
Hint: You can control the top speed of your car by changing the `acceleration` variable in `ControlStatus` from `imitations.py`. But be aware that this also changes the actions you are recording.

d) **Data augmentation.** As discussed in the lecture, the more versatile the training data is, the better generally. Investigate two ways to create more training data with synthetically modified data by augmenting the (observation, action) - pairs the simulator provides. Does the overall performance change?

e) **Fine-tuning.** What other tricks can be used to improve the performance of the network? You could think of trying different network architectures, learning rate adaptation, dropout-, batch- or instance normalization, different optimizers or class imbalance of the training data. Please try at least two ideas, explain your motivation for trying them and whether they improved the result.

1.3 DAGGER Implementation (4+5 Points)

a) The traditional approach to imitation learning ignores the change in distribution and simply trains a policy π that performs well under the distribution of states encountered by the expert d_{π^*}. This can be achieved using any standard supervised learning algorithm. It finds the policy $\hat{\pi}_{sup}$:

$$\hat{\pi}_{sup} = \underset{\pi \in \Pi}{\operatorname{argmin}} \ \mathbb{E}_{s \sim d_{\pi^*}}[\ell(s, \pi)] \tag{1}$$

Prove that by assuming $\ell(s, \pi)$ is the 0-1 loss (or upper bound on the 0-1 loss) implies the following performance guarantee with respect to any task cost function C bounded in $[0, 1]$:

Theorem 1.1 *Let* $\mathbb{E}_{s \sim d_{\pi^*}}[\ell(s, \pi)] = \epsilon$, *then* $J(\pi) \leq J(\pi^*) + T^2\epsilon$.

Please also answer when does DAGGER lead to a performance similar to Behavior Cloning? Are there common scenarios in Urban Driving that could lead to quadratic cost?

b) Implement the DAGGER algorithm. Implement the DAGGER algorithm [3] below and apply it to your dataset. Do you observe an improvement in performance? What scenarios benefit most from DAGGER? Compute and plot the regret, what do you observe? Compare the result and report your findings.

Initialize $\mathcal{D} \leftarrow \varnothing$.

Initialize $\hat{\pi}_1$ to any policy in Π.

FOR $i = 1$ **to** N

Let $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$.

Sample T-step trajectories using π_i.

Get dataset $\mathcal{D}_i = \{(s, \pi^*(s))\}$ of visited states by π_i and actions given by expert.

Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.

Train classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$.

END FOR

Return best $\hat{\pi}_i$ on validation.

DAgger Algorithm.

1.4 Competition (0 Points)

With each exercise sheet you are welcome to participate in a non-graded competition! Run `python3 main.py score` using your current model and submit your score to https://goo.gl/forms/7fckYyCUT6tsnQKs2. You can see the rankings here: https://docs.google.com/spreadsheets/d/1kGo6MLRYmGNAcvjKlNbRU4t7xErKqldGNaLsreK7IgA/edit?usp=sharing. The winners for each of the 3 exercise sessions will be asked to present their approach in the very last lecture.
The evaluation works as follows: the models are tested on a set of validation-tracks. For each track, the reward after 600 frames is used as performance measure. And the mean reward from all validation tracks then forms the overall score. For the final ranking, we will run that evaluation on a set of secret tracks for every submission. The reward is -0.1 every frame and +1000/N for every track tile visited, where N is the total number of tiles in track. Good luck!

1.5 References

[1] https://papers.nips.cc/paper/95-alvinn-an-autonomous-land-vehicle-in-a-neural-network.pdf
[2] https://arxiv.org/pdf/1604.07316.pdf