CSC411H1S Project 4

Hao Hui Tan(999741711, tanstev1)
Kyle Zhou (1000959732, zhoukyle)

March 27, 2018

1. The board is represented by a flat 9-element NumPy tuple. Turn denotes whose turn it is (1 for X, 2 for O). Done denotes whether the game is done (True if game is over, False otherwise.)

   Below is an example of a sample game played against myself.

```
Python 2.7.14 |Anaconda custom (64-bit)| (default, Oct  5 2017, 02:28:52)
[GCC 4.2.1 Compatible Clang 4.0.1 (tags/RELEASE_401/final)] on darwin
env.render()
====
env.step(0)
Out[3]: (array([1, 0, 0, 0, 0, 0, 0, 0, 0]), ’valid’, False)
env.render()
x..
====
env.step(4)
Out[5]: (array([1, 0, 0, 0, 2, 0, 0, 0, 0]), ’valid’, False)
env.render()
x..
.o.
====
env.step(8)
Out[7]: (array([1, 0, 0, 0, 2, 0, 0, 0, 1]), ’valid’, False)
env.render()
x..
.o.
..x
====
env.step(2)
Out[9]: (array([1, 0, 2, 0, 2, 0, 0, 0, 1]), ’valid’, False)
env.render()
x.o
.o.
..x
====
env.step(6)
Out[11]: (array([1, 0, 2, 0, 2, 0, 1, 0, 1]), ’valid’, False)
env.render()
x.o
.o.
x.x
====
env.step(3)
Out[13]: (array([1, 0, 2, 2, 2, 0, 1, 0, 1]), ’valid’, False)
env.render()
x.o
oo.
x.x
====
env.step(7)
Out[15]: (array([1, 0, 2, 2, 2, 0, 1, 1, 1]), ’win’, True)
env.render()
x.o
oo.
xxx
====
env.done
Out[17]: True
env.step(1)
Out[18]: (array([1, 0, 2, 2, 2, 0, 1, 1, 1]), 'done', True)
```

2. (a) The following is the new implemented policy

Listing 1:

```python
class Policy(nn.Module):
    """
    The Tic-Tac-Toe Policy
    """
    def __init__(self, input_size=27, hidden_size=64, output_size=9):
        super(Policy, self).__init__()

        self.linear1 = nn.Linear(input_size, hidden_size)
        self.linear2 = nn.Linear(hidden_size, output_size)

    def forward(self, x):
        x = F.relu(self.linear1(x))
        return F.relu(self.linear2(x))
```

   (b) The 27 dimensions are a flattened encoding of a one-hot encoding of the state of the board. If `.view(3,9)` is applied to the array, the columns would be the one-hot encoding of each cell in the board (starting from the top left, going across each row, and ending in the bottom right).

   If a column contains “1 0 0,” the cell is empty.
   If a column contains “0 1 0,” the cell is occupied by an X.
   If a column contains “0 0 1,” the cell is occupied by an O.

   (c) The value in each dimension means the chance that making the move (e.g. adding an X into that cell) would result in winning the game. This policy is stochastic because it samples the next move from a distribution, rather than following a deterministic algorithm.

3. (a)

   (b)

4. (a)

   (b)

5. (a)

   (b)

   (c)

   (d)

6.

7.

8.