CSC 261 Spring 2018
Lab 11: Decision Tree Analysis
May 2nd 2018

1 Value Iteration

Predict: We used $\gamma = .99$ for generating our predictions and running our algorithm on the 4x3 grid world. There are some states for which the policy is fairly obvious, so we can have a good sense of the relative utilities for nearby states. The most obvious cases for this are when we are in a "tunnel", in which we are simply moving along a sequence of states. A state that is farther along in the tunnel by 1 should be about -.04 * *number of expected tries* it take to move along in the sequence better than the state that comes before it. Because we have probability .8 of making progress each turn, this is about $\frac{.04}{.8} = .05$. This does ignore the discount rate, which should draw the utility of all states a little bit towards 0. So we have this relationship between nearby states. Also, it seems like it will be worth it for all states to walk to a +1 terminal state, rather than throwing themselves into a -1 state in order to stop receiving negative rewards. If the state space were more spread out, this might be a good strategy, but as is the distance is not large enough to encourage this. So even states that are close to a -1 could have utility significantly greater than -1. Fortunately, it's pretty easy to avoid accidentally falling into a -1 terminal state - the bottom left terminal state has a state adjacent to it which could accidentally fall into it, but the other -1 states are such that it's easy to guarantee avoiding them. So we probably will not see a huge dropoff in utility around -1 terminal states, which you could see for other nondeterministic transition functions.

Given that (1) states will gradually increase in utility as we approach a +1 terminal state as well as (2) the fact that many of the -1 terminal states are difficult to "accidentally" fall into, a good policy would be one that just takes the shortest path to a +1 terminal state with minimal regard for -1 terminal states. -1 terminal states would not be completely ignored however, but given that they are difficult to accidentally fall into, not much concern would be given to getting close to them.

Experiment: Below are the results from running our value iteration algorithm on the 16x4 maze provided.

-.328	-.281	-.234	-.196		.238	.298	.352	.407	.352		−1	.930	+1		-3.99
		-.186	-.143		.185			.477		.608		.869			
	-.207	-.143	-.089		.133			.533	.602	.666	.741	.801			+1
−1	-.336		-.033	.024	.075		−1	.498	.551	.602		.754	.808	.869	.930

Reflect: Looking back on the predictions that we made in the first bit of this section there are some areas that we made fairly good estimates. For example, looking at the "tunnels" in the maze and counting the steps to go until a +1 terminal state, we see that the estimated utility of -.04 * *number of expected tries* serves as a fairly good predictor (given a small offset of about .1-.15). As for where our predictions came up short, we underestimated the number and magnitude of the negative utilities of the states towards the left hand side of the maze. Although we speculated that there would be some

negative utilities of states in the top left corner, the actual utilities were much lower in that region than expected.

2 Policy Iteration

Predict: Given the utilities we obtained from value iteration on as well as the actions taken on the 4x3 maze it makes sense that the optimal policy would go from the current state to the neighboring state with the highest utility. This holds in the 4x3 maze example with every action taken. As for general trends in the policy, as noted above, no states are so bad that life is painful enough to head toward a nearby -1 terminal state, and so the general trend of policies we will see is flow towards the nearest +1 state. An interesting question is what policy we will have for the toprightmost state - probably whatever default policy the "randomize policy" part of the algorithm gives it, as it has no way to improve its payoff by switching its action. An interesting case that might come up is in a situation like in the bottom left states - you could imagine that it would be a better policy to move right, rather than up, in the bottom left state in order to avoid risking the -1 terminal state. Based on the numbers, it seems to not be worth the negative reward, but if the bottom left state the payoff were more negative, it might be. The general policy we expect will just be a flow toward +1 terminal states, with the general rule that states will go to the highest utility state that is next to them (more precisely, the action with the highest expected utility, as defined by .8 * the utility of the state resulting from the correct move + .1 * the utility of the each state orthogonal to the correct move).

Experiment: Below are the results from running our policy iteration algorithm on the 16x4 maze provided.

3	3	1	1		3	3	3	1	2		−1	3	+1		0
		1	1		0			1		1		0			
	3	3	1		0			3	3	3	3	0			+1
−1	0		3	3	0		−1	3	3	0		3	3	3	0

Reflect: We got about what we would expect. The top right hand policy was 0, which is obviously consistent with it being random, but it's hard to tell from it whether the analysis above was correct. We did see that it was better to walk all the way to a terminal +1 even when very far away, rather than end your misery at a -1 terminal state. Interesting that the prevalence of +1 terminal states on the right side of the maps makes "right" (aka 3) such a common policy. We generally saw what we expected, which was aided by looking at the utilities from the values part prior to doing the prediction step for policies.