Name: ______________________________

Adult supervision

9. (11 points) Sometimes we can make robust reinforcement-learning algorithms by reducing the problem to supervised learning. Assume:

- The state space is $\mathbb{R}^d$, so in general the same state s may not occur more than once in our data set.
- The action space is $\{0, 1\}$.
- The space of possible rewards is $\{0, 1\}$.
- There is a discount factor γ.

You are given a data set $\mathcal{D}$ of experience interacting with a domain. It contains n tuples, each of the form (s, a, r, s'). Let $\mathcal{D}_0$ be the subset of the data tuples where $a = 0$, and similarly $\mathcal{D}_1$ be the subset of the data tuples where $a = 1$.

Assume you have supervised classification and regression algorithms available to you, so that you can call **classify**(X, Y) or **regress**(X, Y) where X is a matrix of input values and Y is a vector of output values, and get out a hypothesis.

In each of the following questions, we will ask you to construct a call to one of these procedures to produce a Q, V, or π function. In each case, we will ask you to specify:

- Whether it is a regression or classification problem.
- The subset of $\mathcal{D}$ you will use.
- How you will construct a training example (x, y) from an original tuple (s, a, r, s').

For example, if you wanted to train a neural network to take in a state s and predict the expected next state given that you take action 1, then you might do a regression problem using data D_1, by setting $x = s$ and $y = s'$.

(a) Assume horizon $h = 1$. Construct a supervised learning problem to find $Q^1(s, 0)$, that is, the horizon-1 Q value for action 0, as a function of state s.

i. ○ Classification √ **Regression**

ii. ○ $\mathcal{D}$ √ $\mathcal{D}_0$ ○ $\mathcal{D}_1$

iii. x: s

iv. y: r

(b) Still assuming horizon $h = 1$, construct a supervised learning problem to find the optimal policy π^1. Recall that the space of possible rewards is $\{0, 1\}$.

i. √ **Classification** ○ Regression

ii. √ $\mathcal{D}$ ○ $\mathcal{D}_0$ ○ $\mathcal{D}_1$

iii. x: s

iv. y: a **if** $r = 1$ **else** $1 - a$

(c) Now, assume that we have already learned $V^3(s)$, that is, a function that maps a state s into the optimal horizon-three value.

Construct a supervised learning problem to find the optimal horizon 4 Q function for action 0, $Q^4(s, 0)$. You can make calls to V^3.

i. ◯ Classification √ **Regression**

ii. ◯ $\mathcal{D}$ √ $\mathcal{D}_0$ ◯ $\mathcal{D}_1$

iii. x: ___s___

iv. y: ___$r + \gamma V^3(s')$___

(d) Because the state space is continuous, it is difficult to train V^4 without first estimating Q^4, given only our data set and V^3. Explain briefly why.

> **Solution:** For any given s we only know what happens when we take one of the actions, but not the other, since they don't line up, we don't have a way to take the max over actions.