.

Advanced Machine Learning:

Making Rational Decisions

Nando de Freitas

January 12, 2006

Lecture 1 - *Introduction*

OBJECTIVE: In this first lecture, we will introduce the subject of this course: learning to make rational decisions under uncertainty. We begin by stating the problem and establishing the basic vocabulary and notation.

◇ LEARNING AND ACTING

Humans are remarkably good at manipulating their environment. They can probe the environment in order to gather information and increase their knowledge. This knowledge can in turn be used to devise better probing strategies. Humans are also very good at constructing abstract long range plans despite the fact that they live in a world that is fraught with constraints, conflicting goals, partial observability, uncertainty, dynamics and nonlinearity.

Just think of the following examples:

- A child learns by playing and interacting with the environment. Testing boundaries is key to their development.
- A graduate student learns what questions she can ask during class to maximize her knowledge of the subject while avoiding possible embarrassment.
- Politicians learn complex, abstract ways of influencing people by observing their reaction to governmental policies on same-sex-marriage, wars, tax and so on. Despite huge uncertainty due to partial observability and lies, many of them learn to make appropriate temporary concessions that ensure huge future rewards such as being re-elected.

In all the above examples, the agents learn to trade-off immediate rewards with future rewards.

In this course, we will investigate ways in which an agent (decision maker or controller) can learn to make rational decisions in stochastic dynamic environments. This problem is often studied under the names of **reinforcement learning (RL), optimal control** and **sequential decision making**.

In some situations, the agent can observe the **state** $\mathbf{x}_t \in \mathcal{X}$ of the environment and time t. Often, however, the agent will only have access to noisy **observations** $\mathbf{y}_t \in \mathcal{Y}$ of the environment.

The agent aims to maximize a **reward function** $r(\cdot) : \mathcal{X} \times \mathcal{A} \mapsto \mathbb{R}$ by choosing a **policy** $\boldsymbol{\pi} \in \Pi$. A policy is a set of deterministic or stochastic **decision rules** $\boldsymbol{\pi} = (\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \ldots)$. Each rule $\mathbf{d}_t$ maps the state of the system $\mathbf{x}_t$ to an allowable **action** $\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)$, with $\mathcal{A} = \bigcup_{\mathbf{x}} \mathcal{A}(\mathbf{x})$.

The reward function is also known as the **utility function**. Equivalently, in the language of control and statistics, the agent might want to minimize a **cost function** or **loss**.

The key principle of utility is that if an agent prefers situation A to situation B, then it should be the case that $r(A) > r(B)$.

Let and agent with access to evidence $\mathbf{y}_t$ carry out an action $\mathbf{a}_t$. This action causes the stochastic system to transition to a new state $\mathbf{x}_{t+1}$ with probability $p(\mathbf{x}_{t+1}|\mathbf{a}_t, \mathbf{y}_t)$.

Since we don't know what the new state might be, we marginalize over it in order to compute the **expected utility** of action $\mathbf{a}_t$ given the evidence $\mathbf{y}_t$:

$$EU(\mathbf{a}_t|\mathbf{y}_t) = \int r(\mathbf{x}_{t+1}, \mathbf{a}_t)\; p(\mathbf{x}_{t+1}|\mathbf{a}_t, \mathbf{y}_t)\; d\mathbf{x}_{t+1}$$

This expression assumes that the state is continuous. In

discrete systems, we have:

$$EU(\mathbf{a}_t|\mathbf{y}_t) = \sum_{\mathbf{x}_{t+1}} r(\mathbf{x}_{t+1}, \mathbf{a}_t)\; p(\mathbf{x}_{t+1}|\mathbf{a}_t, \mathbf{y}_t)$$

Agents can act **rationally** by maximizing this quantity. Even in a game theoretic setting, each agent acts by maximizing her expected utility. This is known as **best response**.

◇ EXAMPLE: CANCER TREATMENT

A patient can be in two possible states. He might be $'healthy'$ or have $'cancer'$. From population studies, we have estimates of the prevalence of the disease. In particular, we know:

$$p(\mathbf{x} = healthy) = 0.9$$

$$p(\mathbf{x} = cancer) = 0.1$$

Let us assume that we also know the following reward matrix:

	$\mathbf{a} = no\ treatment$	$\mathbf{a} = treatment$
$\mathbf{x} = healthy$	0	-30
$\mathbf{x} = cancer$	-100	-20

We have to decide whether to treat the patient or not.

We can apply the principle of maximum expected utility to evaluate the value of the actions $\mathbf{a} = treatment$ and $\mathbf{a} = no\ treatment$. That is, we compute:

$$EU(\mathbf{a}) = \sum_{\mathbf{x} \in \{healthy, cancer\}} r(\mathbf{x}, \mathbf{a})\ p(\mathbf{x})$$

⋆

$EU(\mathbf{a} = treatment) =$

$EU(\mathbf{a} = no\ treatment) =$

◇ EXAMPLE: VALUE OF INFORMATION

Suppose a girl (Apple) can choose among N prospective boyfriends uniformly at random. Only one of the boys will result in L units of love and the others in 0 units of love. The cost of adjusting to a new boyfriend is L/N units of love.

⋆

The expected return with N candidates is:

The reward is the expected return minus the adjustment cost:

A friend offers to tell her whether Steve is the right one, but asks for $2L/N$ units of love in return. Should she accept this offer? Is knowing whether Steve is the right one worth this much?

To answer these questions, we compute the expected reward of the action "check out Steve". The future state can be either Steve is the right one or not. That is we need to compute:

$$EU(checkout) = \sum_{\mathbf{x}_{t+1} \in \{right, wrong\}} r(\mathbf{x}_{t+1})\ p(\mathbf{x}_{t+1}|checkout)$$

⋆ Case 1: $\mathbf{a}_t = checkout\ \mathbf{x}_{t+1} = right$

Steve is the right one with probability:

The cost of adjusting to Steve is:

If Steve is the right one, Apple's reward is:

⋆ Case 2: $\mathbf{a}_t = checkout\ \mathbf{x}_{t+1} = wrong$

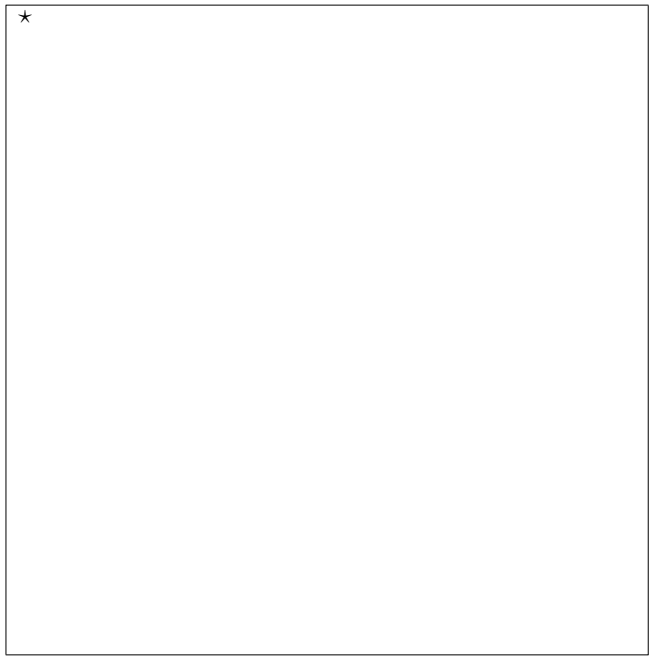
Steve is the wrong one with probability:

If Steve is the wrong one, the probability of finding the right boyfriend becomes:

Even if Steve turns out to be the wrong one, the reward (expected return minus future adjustment cost) for the next stage of the game becomes:

To assess the **value of the information** in units of Love on Steve, we compute the expected utility:

The uncertainty in the new state (and possibly in the new action), results in a distribution over rewards. This distribution can be used to assess the value of information. Consider the following scenarios from the AI book of Stuart Russell:



*"Information has value to the extent that it is likely to cause a change of plan and to the extent that the new plan will be significantly better than the old plan".*S.R.

◇ OPEN AND CLOSED LOOP CONTROL

If Apple's friend had offered her information on Albert as well as on Steve for L/N units each, should she have accepted the offer?

Apple could make a single shot decision. Alternatively, she could first try Steve, observe the new state of the dating game and then decide whether to try Albert. This is the **sequential** approach.

In control theory, the one-shot approach is known as **open loop control** whereas the sequential approach is known as **feedback** or **closed loop control**.

Typically, we need to choose a sequence of decisions while interacting with the environment. In mathematical terms, we must solve the following joint optimization and integration

problem:

$$\max_{\boldsymbol{\pi}} \mathbb{E}\left[r(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1, \mathbf{a}_1, \ldots)\right]$$

where the expectation is taken with respect to the sequence of random variables in the system.

Note that the policy could be a sequence of actions (**pure strategy**) or a sequence of distributions over actions (**mixed strategy**).

The cost of evaluating the reward function is exponential. Typically, one decomposes this function to make the problem tractable:

$$r(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1, \mathbf{a}_1, \ldots) = f\left[r_0(\mathbf{x}_0, \mathbf{a}_0),\ r_1(\mathbf{x}_1, \mathbf{a}_1), \ldots\right]$$

where f is often a simple addition operator.

Designing multi-attribute structured reward functions is an important research problem.

◇ MODEL-FREE AND MODEL-BASED RL

In sequential, model-based RL, the agent chooses action $\mathbf{a}_t \in \mathcal{A}(\mathbf{x}_t)$ in state $\mathbf{x}_t$ (assuming that the state is perfectly observable) and receives a reward $r_t(\mathbf{x}_t, \mathbf{a}_t)$. The environment may then transition to state $\mathbf{x}_{t+1}$ according to the dynamic probabilistic model $p_t(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{a}_t)$. The tuple $(\mathcal{X}, \mathcal{A}(\mathbf{x}_t), p_t(\cdot|\mathbf{x}_t, \mathbf{a}_t), r_t(\mathbf{x}_t, \mathbf{a}_t))$ is known as a **Markov decision process (MDP)**.

If we only have partial observations on the state and an observation model $p(\mathbf{y}_t|\mathbf{x}_t)$, we augment the MDP with this model to obtain a **POMDP**.

POMDPs are harder to solve than MDPs, but then again the world requires the use of POMDPs. Later we will see that POMDPs over world states are MDPs over posterior distributions on the world states.

In model-free RL there is no model describing the probabilistic transitions of the system. We will come back to this type of learning in a few lectures.

The probabilistic dynamic model can be parameterized. The parameters can be learned during the reinforcement stage or during an initial stage of supervised learning.

◇ DECISION RULES AND POLICIES

During a decision instant (**decision epoch**), a **decision rule** specifies a procedure for choosing actions given the state of the system or some information about this state. For now, let us assume that the state is perfectly observable.

Markovian deterministic decision rules are mappings of the form $\mathbf{d}_t : \mathcal{X} \mapsto \mathcal{A}(\mathbf{x}_t)$.

Markovian stochastic (randomized) decision rules are mappings of the form $\mathbf{d}_t : \mathcal{X} \mapsto \mathcal{P}(\mathcal{A}(\mathbf{x}_t))$. That is, they map states to distributions over actions. In game theory, these are known as mixed strategies.

Decision rules can be history dependent. Given the history $\mathbf{h}_t = (\mathbf{x}_1, \mathbf{a}_1, \ldots, \mathbf{x}_{t-1}, \mathbf{a}_{t-1}, \mathbf{x}_t) \in \mathcal{H}_t$, we can have a history dependent deterministic decision rule $\mathbf{d}_t : \mathcal{H}_t \mapsto \mathcal{A}(\mathbf{x}_t)$.

A **policy** is a contingency plan or strategy. It is a sequence of decision rules $\boldsymbol{\pi} = (\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \ldots)$. Recall our problem:

$$\max_{\boldsymbol{\pi}} \mathbb{E}\left[r(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1, \mathbf{a}_1, \ldots)\right]$$

If $\boldsymbol{\pi}$ is stochastic and Markovian, the expectation is taken with respect to $p(\mathbf{x}_0)p(\mathbf{a}_0|\mathbf{x}_0)p(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0)p(\mathbf{a}_1|\mathbf{x}_1)p(\mathbf{x}_2|\mathbf{x}_1, \mathbf{a}_1)\cdots$.

If $\boldsymbol{\pi}$ is deterministic and Markovian, the expectation is taken with respect to $p(\mathbf{x}_0)p(\mathbf{x}_1|\mathbf{x}_0, \mathbf{a}_0)p(\mathbf{x}_2|\mathbf{x}_1, \mathbf{a}_1)\cdots$.

◇ OPTIMALITY CRITERIA

Before moving on to examples, we need to state the optimality criteria that will govern most of these examples. Recall that our original problem is:

$$\max_{\boldsymbol{\pi}} \mathbb{E}\left[r(\mathbf{x}_0, \mathbf{a}_0, \mathbf{x}_1, \mathbf{a}_1, \ldots)\right]$$

A more tractable criterion is the following separable expected reward:

$$\max_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{t=0}^{N-1} r_t(\mathbf{x}_t, \mathbf{a}_t) + r_N(\mathbf{x}_N)\right]$$

where the last reward depends only on the final state. Note that in this case the number of decisions is bounded. This is known as a **finite horizon** problem.

There exists another convention where the reward at time t also depends on the new states. That is, we reward transi-

tions:

$$\max_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{t=0}^{N-1} r_t(\mathbf{x}_t, \mathbf{a}_t, \mathbf{x}_{t+1}) + r_N(\mathbf{x}_N)\right]$$

Sometimes, there is no obvious choice of N. If the system is guaranteed to visit a **terminating state** in finite time and produce zero reward thereafter, one could simply use the following **infinite horizon** cost:

$$\max_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} r_t(\mathbf{x}_t, \mathbf{a}_t)\right]$$

If the system produces ongoing rewards, one can adopt a **discounted objective**

$$\max_{\boldsymbol{\pi}} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(\mathbf{x}_t, \mathbf{a}_t)\right]$$

where $\gamma \in [0, 1)$ is a **discount factor** that weighs present and future costs. Since the rewards at each decision epoch are bounded, the discounted cost is also bounded:

Another popular way of dealing with infinite horizon problems is to adopt an **average cost** criterion:

$$\max_{\boldsymbol{\pi}} \lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{t=0}^{N-1} r_t(\mathbf{x}_t, \mathbf{a}_t)\right]$$

Finally, a simple objective is to choose the action that maximizes the expected cost at each decision epoch:

$$\max_{\mathbf{a}_t} \mathbb{E}_{p(\mathbf{x}_{t+1}|\mathbf{x}_t,\mathbf{a}_t)} \left[r_t(\mathbf{x}_t, \mathbf{a}_t, \mathbf{x}_{t+1}) \right]$$

This criterion fails to take into account future costs. It is no surprise that it is known as **greedy** or **myopic** decision making.