CS497 Week 1 Notes (1/21, 1/23)

Scribe: Matthew Vagnoni

1 Single-Stage Decision Making: Making A Choice About Something

1.1 Overview

A problem exists like whether to bring an umbrella to class today or not. In this series of lectures, we learn how to model such problems. A single stage is not how to make it from point A to point B, but more so deciding if you should go or not. In some cases, no one but the decision maker is involved and other times someone else might be involved (like the weather, e.g. nature).

1.2 Isolated reasoning: choices on our own

We will start with the most basic of situations and everything will build from there. Consider the following scenario:

Scenario 0 (Discrete)

1. *Let U be a set of n choices, actions or inputs:* $\{u_1, u_2, \ldots, u_n\}$.
2. *Let $L : U \to \mathbb{R}$ be a* loss function *or* cost function. *This can also be thought of as a negative reward.*
3. *Select a $u \in U$ that minimizes $L(U)$. That is, select a decision from the set of all possible decisions, which is better then all other choices.*

In the above scenario if nothing ever changed, we would always arrive at the same conclusion. Such potentially predictable decision making is often called *deterministic* or *pure*. This isn't always the case. Sometime it pays to make a *random* (or *mixed*) decision. Deterministic decisions are nice for some problems like taking an umbrella to class, but not always better then randomly deciding. Perhaps the best reason for wanting to input some degree of randomness into the decision is to make it harder to find a pattern and guess how a decision will be made. This often comes up in games.

Scenario 1 (Randomized)

1. $U = \{u_1, u_2, \ldots, u_n\}$
2. u_i *is selected with probability* p_i *or* $p(u_i)$, $\forall u_i \in U$, *where* $\sum_{i=1}^{n} p_i = 1, p_i \geq 0$.
3. $L : U \to \mathbb{R}$
4. u^* *is a randomized strategy* ($u^* = [p_1,\ p_2,\ \ldots,\ p_n]$)
5. $\mathcal{U}$ *is the set of all randomized strategies.*
6. *Select* $u^* \in \mathcal{U}$ *that minimizes* $\mathrm{E}[L] = \sum_{i=1}^{n} L(u_i)p_i$.

Example 1 (Dice Strategy)

1. *Let the input set* $U = \{a, b, c, d, e, f\}$.
2. *The random strategy,* u^**, is to roll a fair six-sided die.*
3. *If the result is 1, choose* a*; if 2, choose* b*; etc.*
4. *Each* u_i *has the same cost.*

Since the die is fair, this corresponds to choosing $p(a) = p(b) = p(c) = p(d) = p(e) = p(f) = 0.167$.

Here rolling a die is a fine strategy because it does a good job minimizing our expected loss. However, a deterministic strategy would also do just as well. If the cost was 2 for u_1 and 3 for everything else, then rolling a die would not be the best u^*. The best strategy, u^*, would be $p_1 = 1, p_2 = p_3 = p_4 = p_5 = p_6 = 0$. Still, this is no better then a deterministic strategy to always choose u_1. In either case, both types of strategies select the action resulting in minimum loss, and thus arrive at the same conclusion! Instead, randomized strategies shine in games, illustrated in the following example.

Example 2 (Matching Pennies)

1. *Two players simultaneously choose H or T.*
2. *If the outcome is HH or TT, then Player 1 pays Player 2 \$1.*
3. *Otherwise Player 2 pays Player 1 \$1.*
4. *Repeat.*

Using the same strategies from the previous example, now result in totally different outcomes! If player1 always chooses H, then player2 can realize this and also always choose heads. However, if Player1 chooses randomly, with $P(H) = P(T) = .5$, at least Player1 can break even.

So far, we have examined scenarios in which there were only a finite number of possible choices. Many problems, however, have a continuum of choices, as does the following:

Scenario 2 (Continuous)

1. *$U \subseteq \mathbb{R}^d$ (usually, U is closed and bounded)*
2. *$L : U \to \mathbb{R}$*
3. *Select $u \in U$ to minimize L*

Example 3 (Continuum of Choices)

1. *$u = [-1, 1] \subseteq \mathbb{R}$ and $L(u) = u^2$ This is a classical optimization problem. To attain minimum cost we choose $u = 0$.*
2. *However, if $u = (0, 1)$ then we have to find the smallest possible value closest to 0. This introduces the two terms* infimum *, or the closest value to lower bound of the set; and* supremum*, or the biggest value closet to the upper bound of the set. Then, we can still say $\inf_{u \in U} L(u) = 0$.*

1.3 Reasoning about uncertainty: beating nature

In this section, we look at *nature* as a special decision maker. Nature is a decision maker that is presumed to be an unreasoning entity. We do not know what decisions nature will make or has made. Some helpful notation:

1. $\theta \in \Theta$ is a particular choice by nature.
2. Θ is a set of choices for nature. Sometimes this is called the parameter space.
3. Discrete case: $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$. Continuous case: $\Theta \subseteq \mathbb{R}^d$.
4. The loss function is $L(u, \theta)$ or $L : U \times \Theta \to \mathbb{R}$. The operator $\cdot \times \cdot$ is the Cartesian product.

Example 4 (Discrete Case) *The cost for each choice is the cell that the particular u_i intersects the particular θ_i. We do not know which θ_i nature will necessarily pick.*

		U		
		u_1	u_2	u_3
Θ	θ_1	1	-1	2
	θ_2	-1	2	-1
	θ_3	0	-2	1

There are lots of different ways of coping with nature. Two of them are:

1. *Nondeterministic: We can't determine what nature will choose. You could do Murphy's Law and assume no matter what you choose, nature will choose the θ_i to produce the worse outcome. So we want to minimize the maximum expected loss for each u_i. In this case, u_1 is the best choice. Its a very secure if pessimistic choice; any nondeterministic choice is a shoot in the dark.*

2. *Probabilistic: Laplace assumed each choice was equally likely. So he would have chosen u_2. However this is a bit naive. While we might not know exactly what nature will do, we can watch nature and collect some statistics on past outcomes. In any case, probabilistic strategies usually rely heavily upon Bayesian analysis. If we knew from before that $P(\theta_1) = .50$, $P(\theta_2) = .10$, and $P(\theta_3) = .40$. Using this knowledge, we would choose u_1 or u_2. The probabilistic strategy depends of the prior observed distribution of Θ.*

Minimum Regret: Minimum regret is like Homer Simpson and he wants to shout 'Doh' the least amount of times for the choices his made. For each cell in a column, take the cell's value minus the best cell's value. In this example, we would choose u_1. Minimum Regret illustrates the point that there is a great deal of metrics that we can use to choose with, besides Minimum Loss. While it is rather arbitrary for the most part, we will stick with the historical use of loss.

Formally, the two scenarios of this section can be defined as follows:

Scenario 3 (Nondeterministic: Minimax Strategy)

1. $U = \{u_1, \ldots, u_n\}$
2. $\Theta = \{\theta_1, \ldots, \theta_m\}$
3. $L : U \times \Theta \to \mathbb{R}$
4. *Choose u to minimize* $\max\limits_{\theta \in \Theta} L(u, \theta)$.

Scenario 4 (Probabilistic: Expected Optimal Strategy)

1. $U = \{u_1, \ldots, u_n\}$

2. $\Theta = \{\theta_1, \ldots, \theta_m\}$

3. $P(\theta)$ *given* $\forall \theta \in \Theta$

4. $L : U \times \Theta \rightarrow \mathbb{R}$

5. $\mathrm{E}^{\theta}[L] = \sum_{\theta \in \Theta} L(u, \theta) P(\theta)$

1.4 Having a single observation: decisions with data

In the section 1.3 we introduced nature, Θ. One of the ways to improve upon making a decision was using a probabilistic strategy to observe nature and thereby collect statistics of prior outcomes. This section extends from that idea of using information to help make a decision. Some new notation follows:

1. Let y be an *observation*, some data, a measurement, or a sensor reading.

2. Let Y be the *observation space*, the set of all possible y.

3. Let $\gamma : Y \rightarrow U$ denote a *decision rule*, strategy, or plan.

4. Either discrete: $Y = y_1, ..., y_n$ or continuous: $Y \subseteq \mathbb{R}^d$

Carrying our two common nondeterministic and probabilistic strategies from the previous section, we have the following:

Scenario 5 (Nondeterministic)

1. $U = \{u_1, \ldots, u_n\}$

2. $\Theta = \{\theta_1, \ldots, \theta_m\}$

3. $Y = \{y_1, \ldots, y_l\}$

4. *Assume there is some* $F(y) \subseteq \Theta$*, which is known for every* $y \in Y$*.*

5. $L : U \times \Theta \rightarrow \mathbb{R}$

6. *Choose* γ *to minimize* $\max_{\theta \in F(y)} L(\gamma(y), \theta)$ *for each* $y \in Y$*.*

Scenario 6 (Probabilistic: Bayesian Decision Theory)

1. $U = \{u_1, \ldots, u_n\}$

2. $\Theta = \{\theta_1, \ldots, \theta_m\}$

3. $Y = \{y_1, \ldots, y_l\}$

4. $P(\theta)$ *given* $\forall\theta \in \Theta$.
5. $P(y|\theta)$ *given* $\forall y \in Y, \forall\theta \in \Theta$
6. $L : U \times \Theta \to \mathbb{R}$
7. *Then Bayes rule yields* $P(\theta|y) = P(y|\theta)P(\theta)/P(y)$, *or more simply* $P(y|\theta) = \sum_{\theta\in\Theta} P(y|\theta)P(\theta)$[1]
8. *Choose* γ *to minimize* $R(\gamma(y)|y) = \sum_{\theta\in\Theta} L(u,\theta)P(\theta|y)$ *for every* $y \in Y$. *This is the* Conditional Bayes Risk.

Extending the former case, we may imagine that we have k observations: $y_1, \ldots, y_k$. Then, $R(u|y_1, \ldots, y_k) = \sum_{\theta\in\Theta} L(u,\theta)P(\theta|y_1, \ldots, y_k)$. If we assume that $P(y_i|\theta)$ is known for each $i \in \{1, \ldots, k\}$ and that conditional independence holds, we have $P(\theta|y_1, \ldots, y_k) = \left(\prod_{i=1}^{k} P(y_i|\theta)\right) P(\theta)/P(y_1, \ldots, y_k)$. We can again ignore the denominator. This shows that multiple observations can be easily handled. The key is figuring out the appropriate $P(\theta|?)$.

Example 5 (Classification)

1. *Let* $\Omega = \{\omega_1, \ldots, \omega_n\}$ *denote a set of* classes.
2. *Let* y *denote a* feature
3. *Let* Y *denote the* feature space. *The feature set* Y *represents useful information that can help us identify which class an object belongs to.*
4. $\Theta = U = \Omega$. *Nature selects a class,* ω, *of objects from* Ω. *We decide the correct class.*

The cost function is:

$$L(u,\theta) = \begin{cases} 0 & \textit{if } u = \theta \textit{ (correct classification)} \\ 1 & \textit{if } u \neq \theta \textit{ (wrong classification)} \end{cases}$$

Using the Bayesian decision strategy, we minimize the probability of making a classification error.

[1] For the purposes of decision-making, $P(y)$ is simply a scaling factor and may be omitted.