15780: GRADUATE AI (SPRING 2018)

Practice Final

April 28, 2019

1 Social Choice: Strategyproofness

Consider the library allocation problem from slides (Social choice II. Slide 9.), where we pick the location to set up a library. For this problem, we will consider the real plane ($\mathbb{R}^2$) as opposed to the real line ($\mathbb{R}$). Recall that each player has a true preference for the location of the library, which we will refer to as a *peak*.

Assume that the utility function of a player whose peak is $x \in \mathbb{R}^2$ is $-d(x, y)$ for a facility located at y, where d denotes Euclidean distance. Given player peaks $x^1, ..., x^n$, consider the mechanism that locates the library at $(\text{med}\{x_1^i\}, \text{med}\{x_2^i\})$. Prove that this mechanism is strategyproof, i.e., player i cannot increase their utility by reporting a peak that is different from x^i, regardless of the reports of other players.

Note: For simplicity, you can assume that the number of voters n is odd.

2 Probabilistic Modeling: MLE and MAP

(a) [4 points] Given a collection of observed (independent) data points $X = \{x^{(1)}, \ldots, x^{(m)}\}$ from a uniform distribution over $[-2\alpha, \alpha]$ (for $\alpha > 0$), derive the maximum likelihood estimator of α, which maximizes the probability of observing X.

(b) [8 points] Given a collection of observed (independent) data points $X = \{x^{(1)}, ...x^{(m)}\}$ from a uniform distribution over $[0, e^{\alpha}]$ where α follows a prior distribution

$$p(\alpha) \propto e^{-\alpha^2},$$

derive the estimator of α that maximizes the posteriori probability $p(\alpha|X)$. **(Hint: use** $p(\alpha|X) \propto p(X|\alpha)p(\alpha)$**).**

3 Game Theory: IESDS

One method of simplifying the search for Nash equilibria is through the iterated elimination of strictly dominated strategies (IESDS). We say that a player's pure strategy s_i' is strictly dominated by another pure s_i if $\forall s_{-i} \in S_{-i},\ u_i(s_i', s_{-i}) < u_i(s_i, s_{-i})$. In other words, s_1 dominates s_2 if, no matter what the other players do, player i always does strictly better by playing s_1 rather than s_2.

IESDS proceeds by repeatedly eliminating one strictly dominated strategy per round, until there are no more dominated strategies to eliminate. For example, IESDS on the following game proceeds as follows.

	North	East	South	West
Top	2,3	1,-1	4,0	3,-3
Middle	7,2	-2,0	5,2	6,7
Bottom	8,2	0,1	6,-1	4,0

- Column eliminates East, as playing North is strictly better.
- Row eliminates Top, as playing either Middle or Bottom is strictly better now that Column has eliminated East.
- Column eliminates South, as playing West is strictly better now.
- No more strategies can be eliminated; this leaves Row: [Middle, Bottom] and Column: [North, West] as the surviving strategies.

Prove the following: If IESDS eliminates all but one of the strategies of each player, then there is a unique Nash equilibrium in the game.

Hints:

- Start by proving that IESDS will never remove an action s_i that appears (with nonzero probability) in any Nash equilibrium.
- Conclude by applying Nash's Theorem: In any (finite) game, there exists at least one (possibly mixed) Nash equilibrium.

4 Adversarial Attacks

Assume we are given a set of m training points $S = \{(x^{(i)}, y^{(i)}) \in \mathbb{R}^D \times \{-1, +1\} \mid i = 1, \dots m\}$. Consider a monotonically decreasing classification loss $\mathcal{L} : \mathbb{R} \to \mathbb{R}$ and a hypothesis function $h_\theta(x) = \theta^T x$ mapping from $\mathbb{R}^D$ to $\mathbb{R}$ for $\theta \in \mathbb{R}^D$.

For this problem, assume that the training data is such that for every i, the first co-ordinate of $x^{(i)}$ equals its label and all other co-ordinates are zero i.e., $x_1^{(i)} = y^{(i)}$, and $x_j^{(i)} = 0$ for $j > 1$. Consider values θ^a and θ^b of the parameter, that perfectly classify the training data:

$$\theta^a = (1, \overbrace{0, 0, \dots, 0}^{D-1 \text{ zeros}})$$
$$\theta^b = (1, 1, 1, \dots, 1).$$

We can see that for all i, $h_{\theta^a}(x^{(i)}) \cdot y^{(i)} = h_{\theta^b}(x^{(i)}) \cdot y^{(i)} = 1$, leading to perfect classification.

(a) [8 points] **Robustness of θ^a to adversarial attacks.** Consider ϵ such that for every sample i, there exists an adversarial perturbation $\Delta^{(i)}$ satisfying $\|\Delta^{(i)}\|_\infty \leq \epsilon$ and $h_{\theta^a}(x^{(i)} + \Delta^{(i)}) \cdot y^{(i)} \leq 0$? Show that the smallest value ϵ can take is 1.

(b) [8 points] **Robustness of θ^b to adversarial attacks.** Consider ϵ such that for every sample i, there exists an adversarial perturbation $\Delta^{(i)}$ satisfying $\|\Delta^{(i)}\|_\infty \leq \epsilon$ and $h_{\theta^b}(x + \Delta^{(i)}) \cdot y^{(i)} \leq 0$. Show that the smallest value ϵ can take is $1/D$.