Selling Data with Multidimensional Heterogeneity

Andy Haupt
(partly joint work with A. Bonatti, S. Smolin, D. Bergemann)
(partly joint work with A. Agarwal)

Munzer Dahleh Group
MIT LIDS

September 30, 2020

Outline

Pricing Data is a Multi-Dimensional Problem

> *[High-quality outcome data] doesn't do you any good unless you have the right phenotypes of patients.*
>
> — *health data professional*

- We have data $(X_{\text{pheno}}, X_{\text{features}}, Y)$
- Data provider can reduce noise on each of the 3 dimensions
- Makes the allocation problem multi-dimensional
- When buyer uses to train ML model, dimensions interact
- What noise level should the data broker offer? At what price?

Pricing Data for a Machine Learning Algorithm

- Buyer faces supervised Learning Problem $(X_{\text{pheno}}, X_{\text{features}}, Y)$
- They would like to train prediction algorithm $f: X_{\text{features}} \mapsto Y$
- Stakeholder is interested in $\mathbb{E}[\ell(f(X_{\text{features}}), Y)|X_{\text{pheno}} = x]$
- Likely varies in x
- Let's make the strong assumption that buyer only cares about the conditional risk vector

$$(\mathbb{E}[\ell(f(X_{\text{features}}), Y)|X_{\text{pheno}} = x])_{x \in \{\text{phenotypes}\}}$$

- Not all conditional risk vectors can be achieved (set $\mathcal{F}$)

Related Literature

Information Design and Sale Bergemann+ AER '18

Multi-Dimensional Monopolist Problem and Auctions Daskalakis+ EMA '17, Kash+ EC '16, Rui '20

Systems for Machine Learning Agarwal+ EC '17, Agarwal+ '19

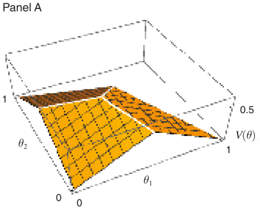

Figure: Berg.+ AER '18

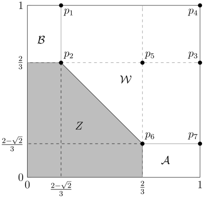

Figure: Das.+ EMA '17

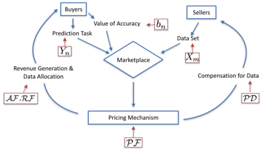

Figure: Agarw.+ EC '17

The Constrained Multi-Goods Monopolist Problem

- Data broker offers menu of quality-price pairs $\mathcal{M} \subseteq \{(q,t) | q \in \mathcal{F}, t \in \mathbb{R}_+\}$, $(0,0) \in \mathcal{M}$, $\mathcal{F} \subseteq \mathbb{R}^n$
- Nature draws a buyer type $\theta \sim F \in \Delta(\mathbb{R}^n)$
- Buyer chooses $(q,t) \in \mathcal{M}$ to maximize $\theta^\intercal q - t$

  Justification for linearity
- Seller wishes to choose $\mathcal{M}$ to maximize $\mathbb{E}_{\theta \sim F}[t(\theta)]$.
- Seller's Problem reduces to LP by the Revelation Principle

  More on Revelation Principle

$$\max_{t(\theta),q(\theta)} \mathbb{E}_{\theta\sim F}[t(\theta)]$$

$$\theta^\intercal q(\theta) - t(\theta) \geq \theta^\intercal q(\theta') - t(\theta'), \quad \theta, \theta' \in \mathbb{R}^n$$

$$\geq 0$$

Selling Noisy Data with

- Data $D = (S_1, X_1, Y_1), (S_2, X_2, Y_2), \ldots, (S_2, X_2, Y_2)$
- S_i is a sensitive attribute
- Seller can offer perturbed data $\tilde{D} = ((\tilde{S}_1, \tilde{X}_1, \tilde{Y}_1), (\tilde{S}_2, \tilde{X}_2, \tilde{Y}_2), \ldots, (\tilde{S}_2, \tilde{X}_2, \tilde{Y}_2))$ at price $p(\tilde{D})$
- Menu $\mathcal{M} = \{(\tilde{D}_i, p(\tilde{D}_i))\}$
- Train predictor w.r.t. commonly known learning technology, e.g. ERM: $f_{\text{ERM}}(\tilde{D}) \in \arg\min_{f \in \mathcal{H}} \mathbb{E}_{(S,X,Y)\sim\tilde{D}}[\ell(f(X), Y)]$
- Evaluate against population distribution

$$-\sum_{s \in S} \theta_S \mathbb{E}_{(S,X,Y)\sim\mathbb{P}}[\ell(f(X), Y) | S = s]$$

Reduction

- Consider the mapping

$$g : \tilde{D} \mapsto (-\mathbb{E}_{(S,X,Y)\sim\mathbb{P}}[\ell(f_{\text{ERM}}(X), Y)|S = s])_{s\in S} \in \mathbb{R}^S$$

- Define the set

$$\mathcal{F} = \{q \in \mathbb{R}^S | \exists \tilde{D} : g(\tilde{D}) = q\} = \text{range}(g)$$

More on the Feasible Payoff Characterization Problem

- Then finding $g(\mathcal{D})$ and t is an instance of the constrained multi-goods monopolist problem

Types of Menus

- $\mathcal{M}$ is grand bundling if $|\mathcal{M}| = 2$.
- $\mathcal{M}$ is upgrade pricing if $\mathcal{M}$ is a chain with respect to the component-wise partial order.
- $\mathcal{M}$ is individual-pricing if for each $(t, q) \in \mathcal{M}$,

$$q = \sum_{i=1}^{n} \lambda_i q_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0 \implies t = \sum_{i=1}^{n} \lambda_i t_i$$

(Informal) Results

Theorem (Informal)

For any full-dimensional $\mathcal{F}$, generically in the type distribution F, individual pricing is suboptimal.

Theorem (Informal)

If the set $\mathcal{F}$ is sufficiently narrow, and the type distribution does not put too much measure on the boundary, then grand bundling is optimal.

Conjecture

For a large class of independent values, there are cases of grand bundling sub-optimality and upgrade pricing optimality.

Formal Definitions

Proof Ideas

1. Reduce to optimization problem for utility function

$$\max \int_\Theta \nabla u(x)^\intercal x - u(x)\,\mathrm{d}F(x),\ u \text{ conv., subgradients in } \mathcal{F}$$

2. Use Green's Theorem to transform to

$$\sup_{u \geq 0} \int u\,\mathrm{d}\mu,\ u \text{ conv., subgradients in } \mathcal{F}$$

3. Use fancy functional analysis to reduce to

$$\inf_\gamma \int_{\Theta^2} \max q^\intercal(\theta - \theta')\,\mathrm{d}\gamma(\theta, \theta'),\ \gamma \text{ coupling, dominance conditions}$$

4. For negative results use necessary conditions for optimality
5. For positive results construct optimal matchings

Last Slide

- We learned:
  1. Data Sale is a problem with multiple dimensions of heterogeneity
  2. We can confidently reduce to the constrained multi-goods monopolist problem
  3. Duality techniques work here to some extent, and allow for both qualitative insights and algorithms
- Discussion Questions:
  - Is revenue-maximization the correct desideratum?
  - In which application domain (online ads, other healthcare) is multidimensional structure particularly well documented?
  - Which data sets could be helpful to study data pricing?

Why linear utility?

- Quasilinear utility in price assumes that money transfers are welfare neutral, i.e. money means the same to rich and poor
- That is often violated
- Linearity here is, however, important to justify pricing we see in the world: We do not offer continuous versions of goods, but post prices
- Linearity in valuation can be justified by assuming that the person has a covariate shift, but evaluates risk.

Go Back

Revelation Principle

- Reduction is based on direct revelation mechanisms
- Instead of choosing a menu item (q, t), we assign type θ a menu item $(q(\theta), t(\theta))$ such that they prefer it to all other assigned $(q(\theta'), t(\theta'))$
- This yields the variational problem

Go Back

Characterizing Feasible Quality Vectors

- Two problems:

  Feasibility Oracle For a $q \in \mathbb{R}^S$, is there a $\tilde{D}$ that produces $g(\tilde{D}) = q$?

  Perturbation Oracle For $q \in \mathbb{R}^S$, compute $\tilde{D}$.
- The former we can formulate as a PAC learning oracle.
- The latter can be approximately in Wasserstein distance with high probability in dataset.

Go Back

Formal Definitions

- $\mathcal{F}$ is narrow if $\max_{q\in\mathcal{F}} \mathbb{1}^\intercal q - \min_{q\in\mathcal{F}} \mathbb{1}^\intercal q$
- $\mathcal{F}$ is called full-dimensional if it cannot be embedded into a lower-dimensional linear supspace
- We call a set of distributions generic if it is open and dense with respect to Wasserstein-2 metric
$\mathcal{W}^2(\mu,\nu) = \sqrt{\inf_\gamma \mathbb{E}_{(X,Y),X\sim\mu,Y\sim\nu}[\|X-Y\|_2^2]}$

Go Back