Selling Data To a Machine Learner: Pricing via Costly Signaling

Junjie Chen[1](speaker), Minming Li[1], Haifeng Xu[2]

[1] City University of Hong Kong
[2] University of Chicago

July 10, 2022

Table of Contents

Introduction

Motivation

One seller sells training data to one buyer, i.e., machine learner who would like to train his ML model

Properties make the sale of data crucially different:

- Ex-ante unverifiable: It is hard to determine beforehand whether the data is truly useful for training an ML model or not, before you train the model.
- Free duplication: Data can be freely duplicated with negligible marginal cost.

Motivation

One seller sells training data to one buyer, i.e., machine learner who would like to train his ML model

Properties make the sale of data crucially different:

- Ex-ante unverifiable: It is hard to determine beforehand whether the data is truly useful for training an ML model or not, before you train the model.
- Free duplication: Data can be freely duplicated with negligible marginal cost.

The usefulness of data (i.e., quality of data) depends on both the ML model design and the data set (e.g., size).

- which are held respectively by both two parties
- no one knows true quality of data in advance because they cannot access both

Model

Timeline of Designed Model

We propose Pricing via Costly Signaling scheme: Costly Signaling + Pricing

- The seller gives a subset $\mathcal{T}$ to the learner for training model.
- The learner trains the ML model and output a preliminary accuracy r_s, which can be simultaneously observed by two parties.
- Both update an estimation of quality of data $Pr(q|r_s, \mathcal{T})$.

Timeline of Designed Model

We propose Pricing via Costly Signaling scheme: Costly Signaling + Pricing

- The seller gives a subset $\mathcal{T}$ to the learner for training model.
- The learner trains the ML model and output a preliminary accuracy r_s, which can be simultaneously observed by two parties.
- Both update an estimation of quality of data $Pr(q|r_s, \mathcal{T})$.
- After sharing set $\mathcal{T}$ of data points, the seller will immediately lose some amount of sales value, due to the free duplication property. The remaining utility is

$$G(r_s, \mathcal{T}, b) = E[u(r_m, b)|r_s, \mathcal{T}, \mathcal{D}] - u(r_s, b)$$

- Based the r_s and $\mathcal{T}$, the seller uses a posted price mechanism for pricing the remaining data

$$\sum_{r_s} Pr(r_s|\mathcal{T}) \sum_{b} \phi(b) \cdot p_{r_s,\mathcal{T}} \cdot \mathbf{1}\Big\{G(r_s, \mathcal{T}, b) \geq p_{r_s,\mathcal{T}}\Big\}$$

Objective

We maximize the seller's revenue:

- determine the shared subset $\mathcal{T}$ in costly signaling step
- determine the price $p_{r_s,\mathcal{T}}$ for remaining data

Homogeneous Data

Selling Homogeneous Data

We first consider selling homogeneous data, in which each data point contributes equally to the performance of the ML model.

- Only the quantity of data points, $t = \frac{|\mathcal{T}|}{|\mathcal{D}|}$, matters $\lambda(r|q, \mathcal{T}) = \lambda(r|q, t)$
- The optimal mechanism is solved in polynomial time by simple enumeration.

Main Results

We have following main results:

Theorem

If valuation function $u(r, b) = f(r)g(b)$ is multiplicatively separable (e.g. $u(r, b) = r \times b$), then the best strategy for the seller is to sell the entire dataset directly without sharing any data. It also holds for additively separable function (i.e., $u(r, b) = f(r) + g(b)$) if $\lambda(0|q, t > 0) = 0$.

- It suggests that it is not always profitable to share data with the learner in the costly signaling step.

Main Results

We also consider the case where the learner and seller hold different prior over the quality q:

$$\forall q,\ |\mu_{sl}(q) - \mu_{ml}(q)| \leq \epsilon \mu_{sl}(q)$$

Theorem

Let $\epsilon \in [0, 1]$ satisfy constraint above. Then exists a mechanism, whose revenue is at least $OPT - \frac{4\epsilon}{1-\epsilon^2}\bar{u}$, where OPT is the optimal expected revenue under the true prior belief of machine learner and $\bar{u} = \max_{r,b} u(r, b)$.

- It suggests that even though the machine learner's true prior belief is private, the revenue that the seller can obtain is still close to that obtained by knowing the true machine learner's prior, as long as their beliefs do not differ much, i.e., ϵ is small.

Heterogeneous Data

Selling Heterogeneous Data

We model heterogeneous data by selling features:

- Each data point is of M dimensions, i.e., M features which are heterogeneous.

Selling Heterogeneous Data

We model heterogeneous data by selling features:

- Each data point is of M dimensions, i.e., M features which are heterogeneous.

Suppose there M features for sale

- We use a quality vector $q = [q_1, q_2, \ldots, q_M]$ to describe qualities of features
- Given any subset of features $\mathcal{T} = \{i_1, i_2, \ldots, i_k\}$, we use $q_{\mathcal{T}} = [q_{i_1}, q_{i_2}, \ldots, q_{i_k}]$ as the corresponding quality subvector.
- We assume here $\lambda(r|q, \mathcal{T})$ is a point distribution, (f is some function)

$$\lambda(r|q, \mathcal{T}) = \lambda(r|q_{\mathcal{T}}) = \begin{cases} 1, & r = f(q_{\mathcal{T}}) \\ 0, & otherwise \end{cases}$$

Main Results

We apply an limited communication constraint to the subset $|\mathcal{T}| \leq T$.

We first have the following hardness result

Theorem

It is NP-hard to achieve $\frac{e}{e+1} + o(1)$ *approximation to the optimal mechanism.*

Main Results

We apply an limited communication constraint to the subset $|\mathcal{T}| \leq T$.

We first have the following hardness result

Theorem

It is NP-hard to achieve $\frac{e}{e+1} + o(1)$ approximation to the optimal mechanism.

We provide a simple approximation algorithm

Theorem

Oblivious to the choice of $f(q)$, by selling the entire dataset directly without sharing any feature, at least $\frac{1}{k}OPT$ revenue can be obtained, where k is the number of private types and OPT is the optimal revenue.

- Note that when $k = 2$, the approximation ratio 0.5 is closed to $\frac{e}{e+1} \approx 0.7$

Conclusions

Conclusions

We take the first step to consider the problem of selling data to a machine learner. A *Pricing via Costly Signaling* scheme is proposed for the homogeneous data and heterogeneous data cases.

Open problems left:

- Is there a better approximation algorithm for approximating the heterogeneous case problem?
- Besides selling data by pricing via costly signaling, is there a better mechanism to sell data under the same setting?