社会计算
Social Computing

数据定价

赵明昊 树扬

mhzhao@dase.ecnu.edu.cn

yshu@dase.ecnu.edu.cn

華東師範大學
EAST CHINA NORMAL UNIVERSITY

Acknowledgement

- 数据定价与交易研究综述，by 江东, 袁野, 张小伟, 王国仁
- A Survey on Data Pricing: from Economics to Data Science, by Jian Pei
- What is Your Data Worth? Quantifying Data Value in Machine Learning, by Ruoxi Jia

课堂内容

- 数据定价的概念
- 数据定价的准则
- 数据定价的方法
- 具体案例：机器学习中的数据定价

数据交易-对象

- **数据拥有者**：向数据平台提供数据，并接受数据平台给予的相应补偿
- **数据平台**：收购数据，集成整合，设定数据收购价格并补偿数据拥有者，设定数据出售价格，为数据消费者提供查询其希望购买数据的接口和服务，给数据消费者提供数据并对出售的数据提供隐私、版权保护等任务
- **数据消费者**：向数据平台提出需求，并支付金钱从数据平台购买所需数据

数据定价与交易研究综述，江东，袁野，张小伟，王国仁

数据交易-生命周期

- **数据收集与集成**：解决数据“从无到有”的问题, 对源数据执行整合、清洗和验证等操作, 以便满足后续数据管理要求以及数据消费者的数据查询要求;
- **数据管理与分析**：解决数据组织存储形式的问题, 同时对数据进行分析以得到其适用范围、出售模式和近似商业价值;
- **数据定价**：关注各种确定数据价格的方法;
- **数据交易**：重点考虑数据市场类型、参与交易各方行为等对数据出售价格的影响

数据定价 What Is Pricing?

- The practice that a business sets a price at which a product or a service can be sold
- Often part of the marketing plan of a business
- Objectives in pricing
  - Profitability
  - Fitness in marketplace
  - Market positioning
  - Price consistency across categories and products
  - Meeting or preventing competitions
  - ...

A Survey on Data Pricing: from Economics to Data Science, Jian Pei

数据定价的特殊性

Cost reduction is the core in production, distribution and consumption of information goods comparing to physical products

- **Search** costs: platform, online search, rare and long tail products
- **Production** costs: innovative business, query-based data consumption
- **Replication** costs: nonrival, bundling, making public (open source), copyright
- **Transportation** costs: Internet
- **Tracking and verification**: track users, personalized markets, price discrimination, privacy sensitive and hold the information

Digital and data economics: investigation of how standard economic models adjust when major costs are reduced dramatically

版本控制（versioning）

- **Linking Price to Value:** setting the price reflecting the value that a customer places on the information
- **Versioning strategy:** making different versions to appeal to different types of customers；Examples: software, movies, ...
  - Delay in information delivery
  - Access convenience
  - Comprehensiveness of information
  - Information manipulation
  - User community
  - Annoyance
  - Customer support
  - ...
- Most versions of information goods are created by subtracting value from the most technologically advanced and complete version

数据定价与交易的准则

- 收入最大化（revenue maximization）
- 真实性（truthful）
- 公平性（fairness）
- 无套利（arbitrage free）
- 隐私保护（privacy preservation）
- 计算高效（computational efficiency）
- ……

收入最大化（revenue maximization）

- **原因与依据：** for a business to be successful long term, a more immediate and important requirement is to win over as many customers as possible
- 数据产品的特殊性：在竞争市场中, 传统产品在边际成本等于边际收益时达到卖家收入最大化，然而数据产品边际成本几乎为0。
- Three streams of revenues for digital products that are delivered online
  - **Money:** a provider can sell to customers content, or more broadly services, such as movies and e-books
  - **Information/privacy:** a provider can collect customer information by tracking (e.g., using cookies) and sell the information about customers to generate revenues
  - **Time/attention:** a provider can sell space in their digital products to advertisers to produce revenue

真实性（truthful）

- 买卖双方都是利己的，并且仅提供能够使得自己利益最大化的价格。
- 如果买家认定了某个产品的购买价格，他就不会再支付多于该价格的金钱去购买此产品。
- 无论其他人如何操作，对于任意的供应商和消费者来说，都不能通过虚报真实价值来增加其收益。
- 真实性保证了每个人在参与交易时都不进行虚假操作。
- 防止数据交易时的猜测行为，并减少市场竞标策略中不必要的开销。

公平性（fairness）

- 数据通常来自不同贡献者，为了确保卖家出售数据的积极性, 数据交易平台需要保证总收入在所有数据贡献者中按其贡献公平分布。
- **沙普利值 (Shapley value)**

$$s_i = \frac{1}{N} \sum_{S \subseteq \Delta \setminus z_i} \frac{1}{\binom{N-1}{|S|}} [v(S \cup \{z_i\}) - v(S)] \tag{2}$$

其中, N 代表博弈参与者, S 表示参与者组成的任意联盟, $v(S)$ 表示联盟 S 的收益函数, z_i 表示联盟 S 中的某个参与者. 沙普利的实现满足如下要求.

(1) 集体理性 (group rationality): 交易获得的收入必须全部分配给所有卖家.

(2) 公平性 (fairness): 对于一个卖家联盟 S 和另外两个卖家 s 和 s', $s, s' \notin S$, 若 $S \cup \{s\}$和 $S \cup \{s'\}$获得了相同的资金, 那么 s 和 s'也应该收到相同的回报. 即, 对于效用的贡献度相同的卖家, 他们所收到的回报也应该相同; 对于一个卖家联盟 S 和一个额外的卖家 $s \notin S$, 若 $S \cup \{s\}$和 S 获得了相同的资金, 则 s 收到的回报为 0. 即, 没有贡献就没有回报.

(3) 可加性 (additivity): 如果分别为两个任务 T_1 和 T_2 回报 v_1 和 v_2, 那么完成两个任务 T_1+T_2 的回报是 v_1+v_2.

无套利（arbitrage free）

- **套利**：买家通过某种手段, 按照低于卖家规定价格获取数据产品。
- Example: an article price: $35, monthly subscription rate $25
- 套利机会的存在会导致数据定价的不一致性, 并使得信息泄露的风险大大增加。

无套利（arbitrage free）

- **多账户套利:** 对于查询束 Q = Q 1 + Q 2 , 买家分别创建两个账号去购买查询 Q 1 和 Q 2 , 平台应该保证查询 Q 的价格至多不能大于查询 Q 1 和 Q 2 价格之和
- **后处理套利:** 如果查询束 Q 1 得到信息是查询束 Q 2 得到信息的子集, 那么 Q 1 价格必须低于 Q 2 价格
- **偶然套利:** 如果买家希望随机购买符合要求的任意一条数据, 如果随机查询的价格低于确定查询某条记录的价格时, 会发生偶然套利
- **确定套利:** 在平台给买家返回有噪声的查询结果并以噪声程度确定查询价格的模式下, 当平台向查询结果内添加的噪声具有较大方差时, 返回结果的精确度也会存在接近 1 的情况, 如果平台给具有较大方差的结果设置较低价格, 则会出现套利情况

数据定价方法

- **基于任务的定价:** 依据该数据产品对于数据消费者执行某项任务所能产生的价值来确定价格（基于查询的定价、基于模型的定价）

- **基于价值的定价:** 依据该数据产品的内在价值如隐私包含程度、数据质量优劣来确定其价格（基于隐私补偿的定价、基于数据质量的定价）

- **基于经济学的定价:** 在确定基础价值的前提下, 依靠市场如供需关系、市场类型等经济学方法来确定数据价格, 主要考虑市场类型和参与人行为对价格产生的影响 （基于花费的定价、基于供需关系的定价、基于博弈论的定价、基于拍卖的定价）

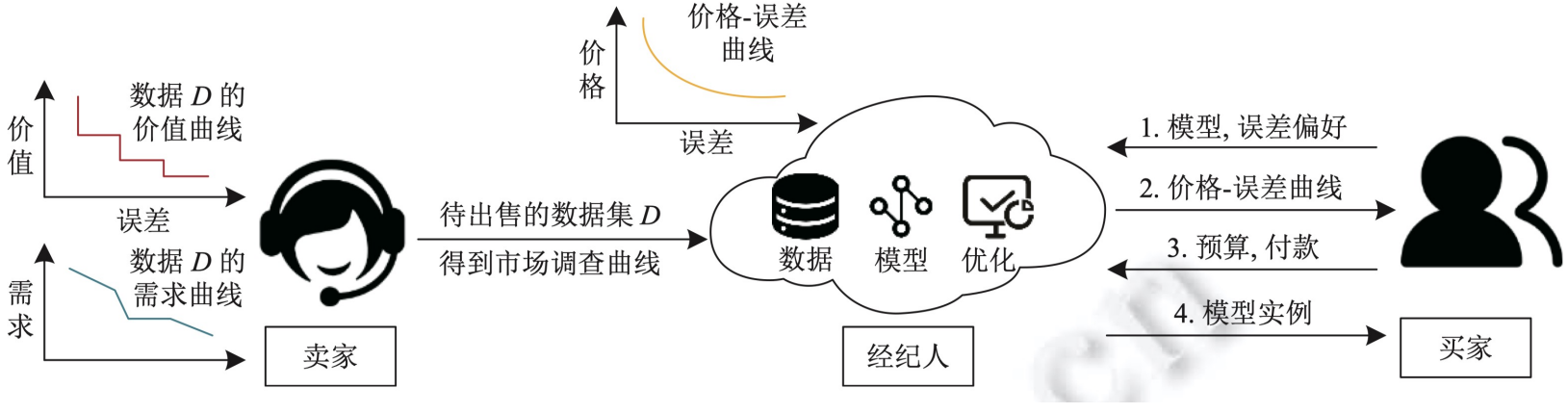
基于查询的定价

- 数据的不同视图（view）作为不同的版本（version）
- 卖家给一定数量的数据视图指定价格，当查询来临时，根据与查询结果相关的视图价格设定查询价格（例如，设定为与查询结果相关的所有视图价格和的最小值）

The key idea in Qirana is that it regards a query as an uncertainty reduction mechanism. Initially, a buyer faces a set of possible databases $\mathcal{I}$ defined by a database schema, primary keys and predefined constraints. Once a buyer obtains the answer E to a query Q, all possible databases D such that $E \neq Q(D)$ are eliminated. The price assigned to Q should be a function of how much the set of possible databases shrinks. Let $\mathcal{S}$ be the set of possible databases before the query Q is answered. $\mathcal{S}$ is called the support set. Then, a weighted coverage function assigns a weight w_i to every $D_i \in \mathcal{S}$, and computes the price to a query by $p^{wc}(Q, D) = \sum_{Q(D_i) \neq Q(D)} w_i$.

基于模型的定价

- 衡量数据点对机器学习模型的贡献度，以此为依据对其进行定价
- 也可为训练好的机器学习模型实例定价
- 可为买家提供不同版本模型（噪声注入），以此进行定价



基于数据质量的定价

- 从不同维度对数据质量进行度量评分和版本控制

$$\text{Value of data} = \text{fixed cost} + \sum_i w_i \cdot \text{factor}_i$$

- 从自身角度考虑数据价值以及从消费者角度考虑数据效用, 通常具有较好的透明度以及较高的可解释性

基于隐私补偿的定价

- 隐私补偿：应对卖家在数据交易中产生的隐私损失问题，激励更多人出售个人数据

- 如何衡量隐私损失，差分隐私

假设数据集 T 经随机算法 M 处理后的输出结果集合为 Y, Y 的任意子集为 D, 对于任意邻近数据集 T 和 T', 若算法满足不等式:

$$\frac{\Pr(M(T)=D)}{\Pr(M(T')=D)} \leqslant e^{\varepsilon} \tag{3}$$

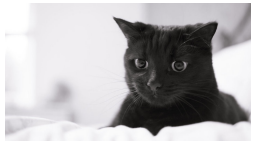
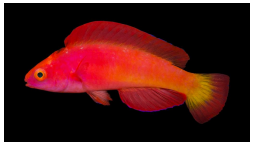
则称算法 M 提供了 ε 差分隐私保护. 差分隐私的出现, 解决了数据定价中衡量隐私损失的问题, 现阶段基于隐私补偿的定价方法大都采用了差分隐私的 ε 值作为确定价格的参数.

基于经济学的定价

- 成本与利润：收集、存储、复制成本; 市场供需关系
- 基于博弈论的定价; 基于拍卖的定价（密封拍卖、双边拍卖）
  - Ascending-bid auction (aka English auction): the price is raised successively until only one bidder remains, who wins the object at the final price
  - Descending auction (aka the Dutch auction): start at a very high price and lower the price continuously, until the first bidder calls out and accepts the current price
  - First-price sealed-bid auction: every bidder submits a bid without knowing the others' bids, and the one making the highest bid wins and pays at the named price
  - Second-price sealed-bid auction (aka the Vickrey auction): every bidder submits a bid without knowing the others' bids, and the one making the highest bid wins and pays only the second highest bid

机器学习中的数据定价

The Shapley value

z_1 s_1

z_2 s_2

z_N s_N

D

$$s_1 + s_2 + \cdots + s_N = U(D)$$

Definition

$$s_i = \frac{1}{N} \sum_{S \subseteq D \setminus \{z_i\}} \frac{1}{\binom{N-1}{|S|}} \left[U(S \cup \{z_i\}) - U(S) \right]$$

#/training pts

Marginal contribution of z_i

Utility function

S is a subset of pts except z_i

$D = \{z_1, \ldots, z_N\}$ is the training set

Properties

Equitable:

- If $U(S \cup \{i\}) = U(S \cup \{j\})$ for all $S \subseteq D : s_i = s_j$
- If $U(S \cup \{i\}) = U(S)$ for all $S \subseteq D : s_i = 0$

Cumulative: $s_i^{U+V} = s_i^U + s_i^V$

Tutorial: What is Your Data Worth? Quantifying Data Value in Machine Learning, Ruoxi Jia

机器学习中的数据定价

Computational challenges

Exponential complexity

Evaluating $U(S)$ in ML is expensive

$$s_i = \frac{1}{N} \sum_{S \subseteq D \setminus \{z_i\}} \frac{1}{\binom{N-1}{|S|}} \Big[U(S \cup \{z_i\}) - U(S) \Big]$$

Monte Carlo approx.

$$= \frac{1}{N!} \sum_{\pi \in \Pi(I)} [U(P_i^\pi \cup \{z_i\}) - U(P_i^\pi)]$$

Set preceding z_i in permutation π

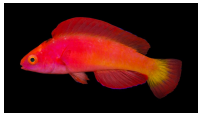
$$= E_\pi \big[U(P_i^\pi \cup \{z_i\}) - U(P_i^\pi) \big]$$

$$\approx \frac{1}{M} \sum_{m=1}^{M} \big[U\big(P_i^{\pi_m} \cup \{z_i\}\big) - U\big(P_i^{\pi_m}\big) \big]$$

Existing work
[Mann et al. 1960, Castro et al. 2008, Maleki et al. 2015, Agarwal et al. 2019]

机器学习中的数据定价

How to Better Make Use of a Single Utility Evaluation?

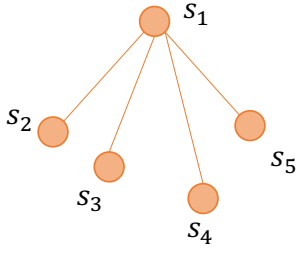
$U($    $)$

$\beta_1 = 1$ $\beta_2 = 1$ $\beta_3 = 1$ $\beta_4 = 0$ $\beta_5 = 0$

Intuition: if the utility on this subset is high, all the points **present** in the subset should have **high** value, all points **absent** should have **low** value

Smartly design the sampling distribution $\beta_1, \ldots, \beta_5 \sim q(\beta)$

$$E_\beta[(\beta_2-\beta_1)U(\beta_1,\ldots,\beta_5)] = s_2 - s_1$$
$$E_\beta[(\beta_3-\beta_1)U(\beta_1,\ldots,\beta_5)] = s_3 - s_1$$
$$E_\beta[(\beta_4-\beta_1)U(\beta_1,\ldots,\beta_5)] = s_4 - s_1$$
$$E_\beta[(\beta_5-\beta_1)U(\beta_1,\ldots,\beta_5)] = s_5 - s_1$$



Towards Efficient Data Valuation Based on the Shapley Value, Ruoxi Jia

谢谢！

華東師範大學
EAST CHINA NORMAL UNIVERSITY