Data valuation by Peer Prediction

Part of AAAI-23 tutorial
The Economics of Data and Machine Learning

Shuran Zheng, February 2023

Economics of data

- How do we price/evaluate a dataset (for a Machine Learning problem)
- How self-interested agents will respond to the pricing/data valuation metric

Motivation

Self-interested data providers

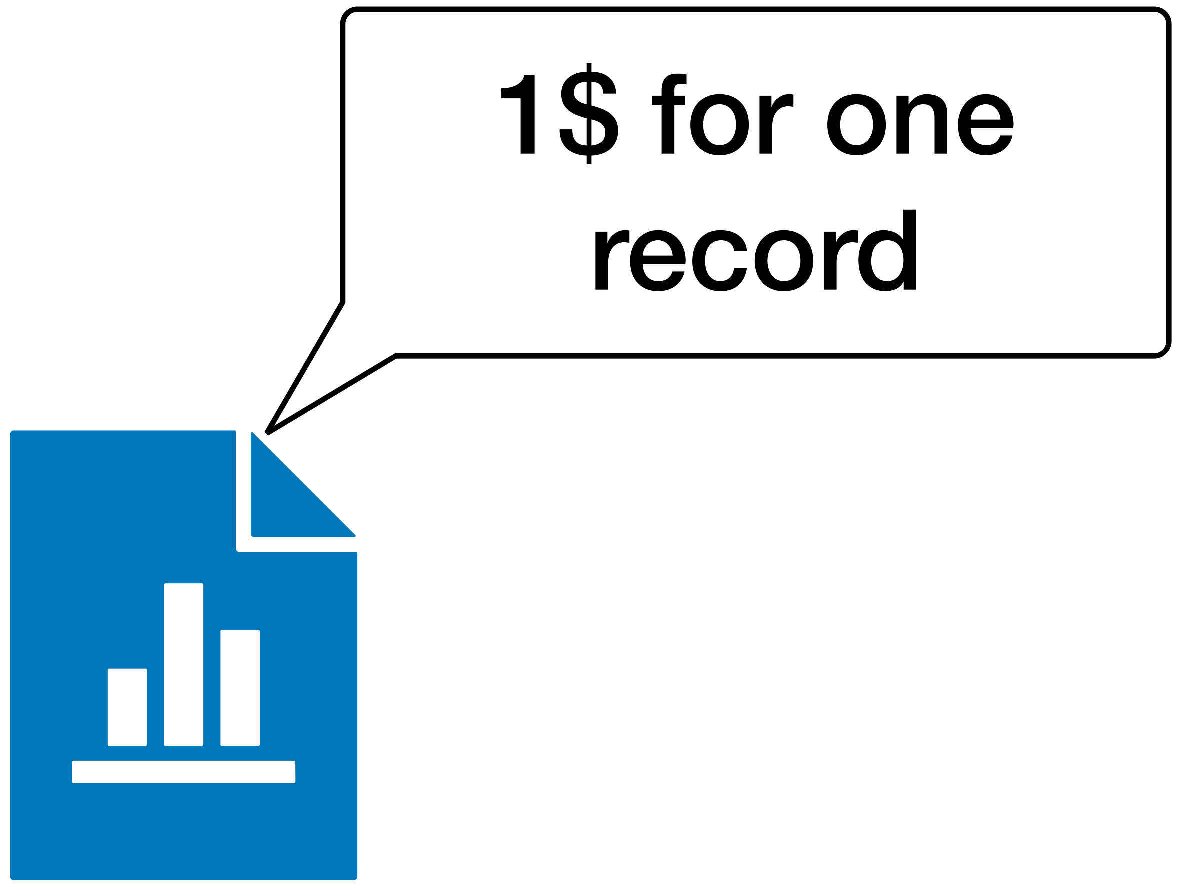
Data providers respond to the data valuation method **strategically**: they respond in a way that maximizes **their own reward**

Motivation

Self-interested data providers

Data providers respond to the data valuation method **strategically**: they respond in a way that maximizes **their own reward**

- E.g. reward data provider proportional to the size of dataset

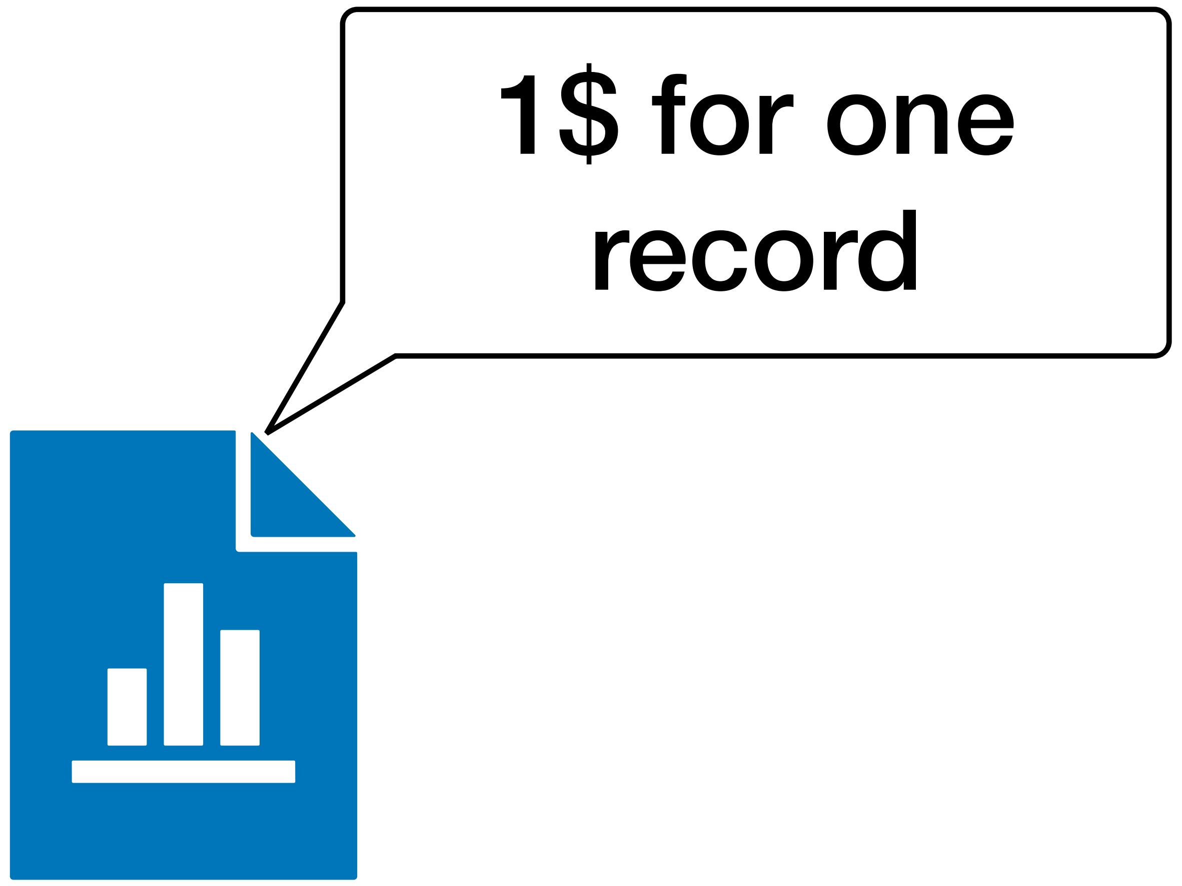


Motivation

Self-interested data providers

Data providers respond to the data valuation method **strategically**: they respond in a way that maximizes **their own reward**

- E.g. reward data provider proportional to the size of dataset
    - duplicate their data

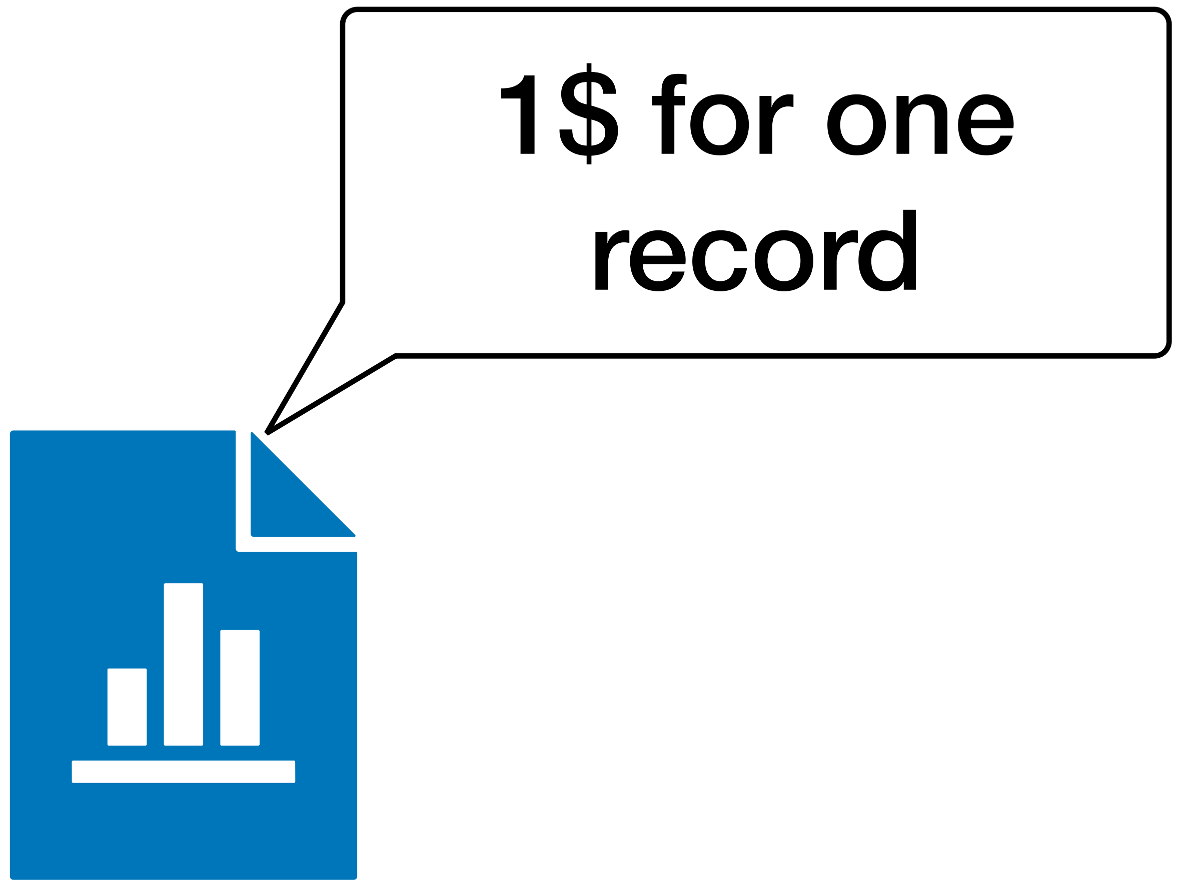


Motivation

Self-interested data providers

Data providers respond to the data valuation method **strategically**: they respond in a way that maximizes **their own reward**

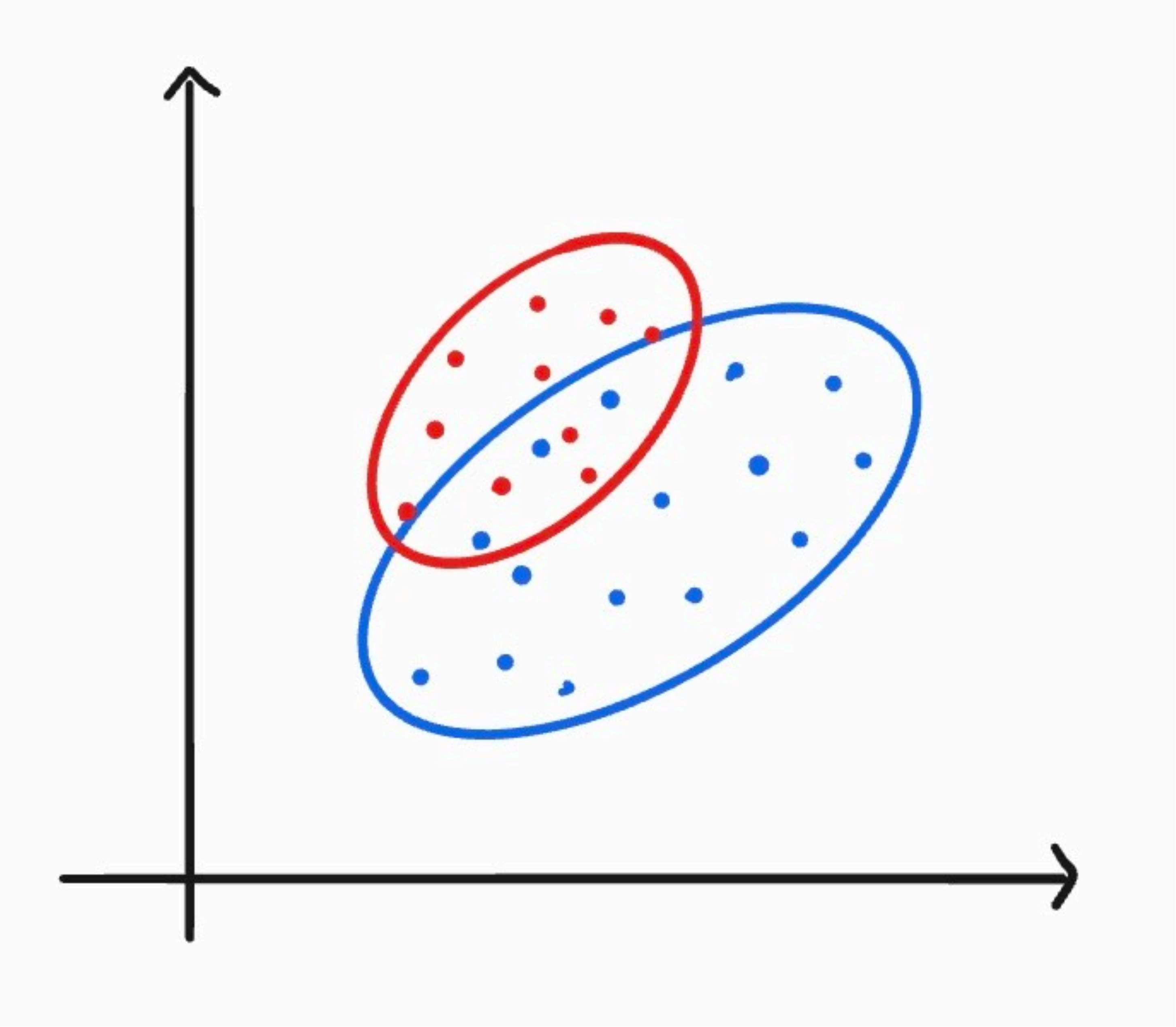
- E.g. reward data provider proportional to the size of dataset
  - duplicate their data
  - generate random data



Motivation

Self-interested data providers

- **E.g. use test data:** train a model on the provided data, reward data provider according to the performance of the model on a test dataset
- Provide data that "matches" the test data
- My data: 50% red, 50% blue
- Test data: 1% red, 99% blue
- Better off dropping some red data



Goal of the talk

A data valuation method that prevents data manipulations

- A data provider holds an original/authentic dataset D
- Any manipulation on the data: **NO**
  - Manipulation on a dataset D: **apply a function** on the dataset $f(D) = D'$
  - Append fake data, duplicate, deletion…

Outline

A data valuation method that prevents data manipulations

- Bayesian modeling & the log scoring rule
- Computing the log scoring rule for Bayesian machine learning
- Sensitivity analysis
- Summary & extension

Outline

A data valuation method that prevents data manipulations

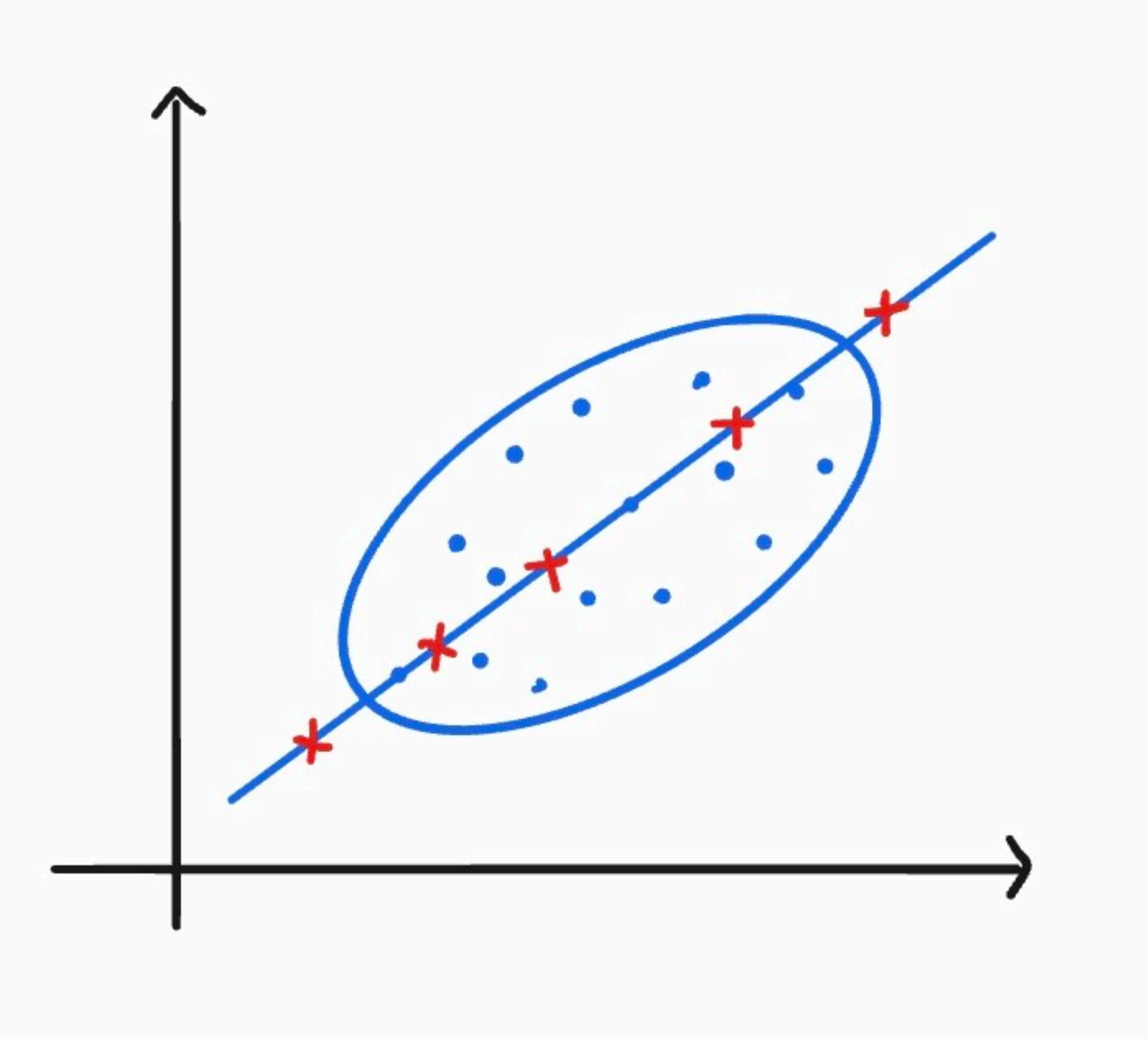
- Bayesian modeling & the log scoring rule
- Computing the log scoring rule for Bayesian machine learning
- Sensitivity analysis
- Summary & extension

Bayesian modeling

Warm up

<u>**Simple observation:**</u> it is not possible to prevent data manipulation if there is **no uncertainty** in the best model on the **test data**

- Know that the test data gives θ^*
- Submit data that gives θ^*



Bayesian modeling

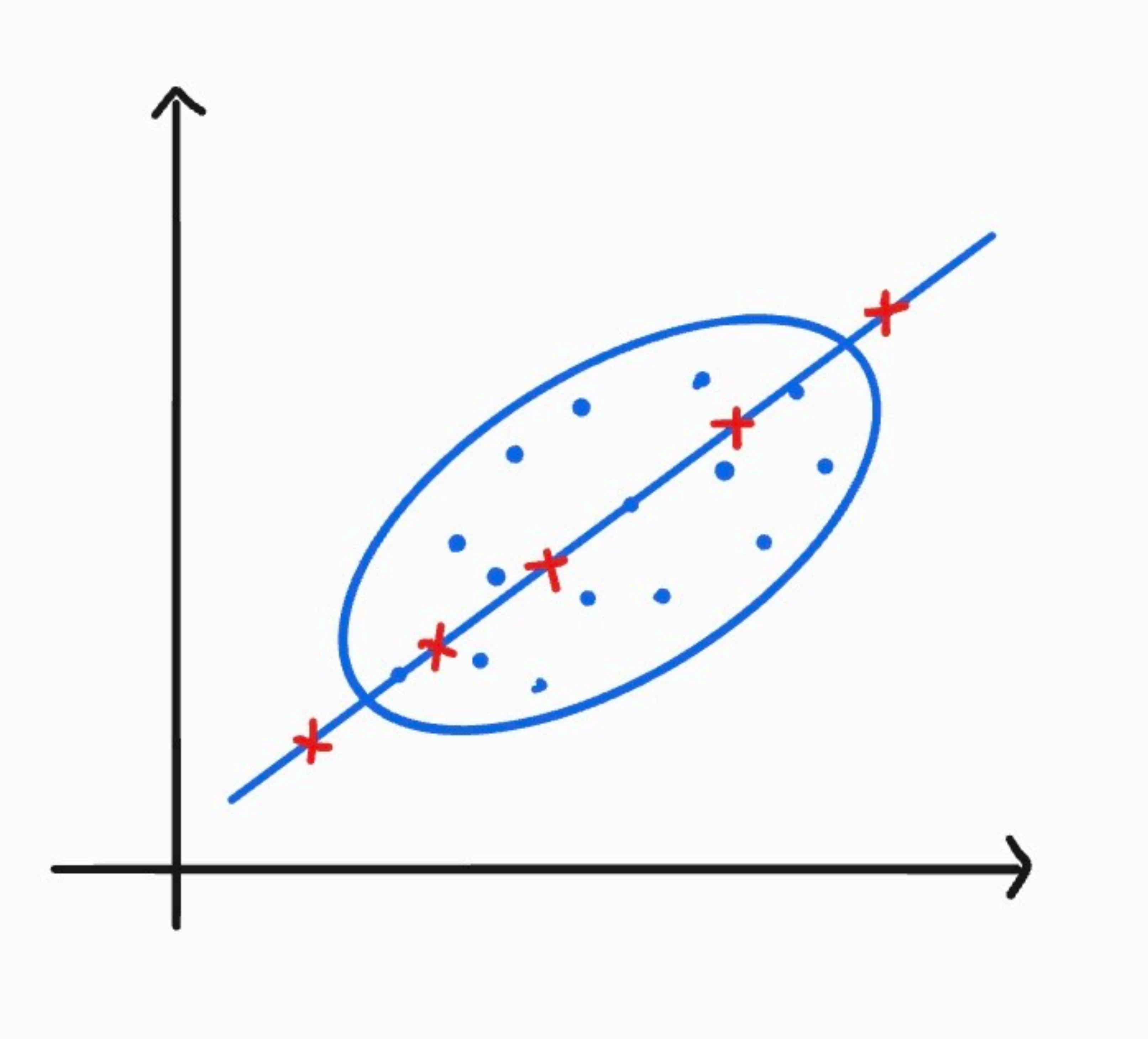
Warm up

<u>**Simple observation:**</u> it is not possible to prevent data manipulation if there is **no uncertainty** in the best model on the **test data**

- Know that the test data gives θ^*
- Submit data that gives θ^*

Uncertainty about θ

=> **Bayesian modeling**



Bayesian modeling

Warm up

A weighted coin with probability of head θ

- Prior about θ: **highly likely θ is large**

Bayesian modeling

Warm up

A weighted coin with probability of head θ

- Prior of θ: **highly likely θ is large**

$$P(\theta = 0.8) = 0.9$$

$$P(\theta = 0.2) = 0.1$$

Bayesian modeling

Warm up

A weighted coin with probability of head θ

- Prior of θ: **highly likely θ is large**

$P(\theta = 0.8) = 0.9$

$P(\theta = 0.2) = 0.1$

- Collect a coin flip X from a data provider
- Test data: a coin flip Y
- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

Bayesian modeling
Warm up

A weighted coin with probability of head θ

- Prior of θ: **highly likely θ is large**

$P(\theta = 0.8) = 0.9$

$P(\theta = 0.2) = 0.1$

- Collect a coin flip X from a data provider
- Test data: a coin flip Y
- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

- Maximize my expected reward
- Best strategy?

Bayesian modeling

Warm up

A weighted coin with probability of head θ

- Prior of θ: **highly likely θ is large**

$P(\theta = 0.8) = 0.9$

$P(\theta = 0.2) = 0.1$

- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

Flips the coin, sees a head $X = H$

- Report $X' =$ **?**

Flips the coin, sees a tail $X = T$

- Report $X' =$ **?**

Bayesian modeling

Warm up

The data provider's strategy depends on her belief about Y, that is, $P(Y = H \mid X)$

- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

Bayesian modeling

Warm up

The data provider's strategy depends on her belief about Y, that is, $P(Y = H \mid X)$

- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

Flips the coin, sees a head $X = H$

- $\Pr(Y = H \mid X = H)$**?**

Flips the coin, sees a tail $X = T$

- $\Pr(Y = H \mid X = T)$**?**

Bayesian modeling

Warm up

The data provider's strategy depends on her belief about Y, that is, $P(Y = H \mid X)$

- How to compute $P(Y = H \mid X)$?
- Based on $P(\theta \mid X)$

- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

Flips the coin, sees a head $X = H$

- $\Pr(Y = H \mid X = H)$**?**

Flips the coin, sees a tail $X = T$

- $\Pr(Y = H \mid X = T)$**?**

Bayesian modeling

Warm up

The data provider's strategy depends on her belief about Y, that is, $P(Y = H \mid X)$

- How to compute $P(Y = H \mid X)$?
- Based on $P(\theta \mid X)$

$P(\theta = 0.8 \mid X)$

$P(\theta = 0.2 \mid X)$

- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

Flips the coin, sees a head $X = H$

- $\Pr(Y = H \mid X = H)$**?**

Flips the coin, sees a tail $X = T$

- $\Pr(Y = H \mid X = T)$**?**

Bayesian modeling

Warm up

The data provider's strategy depends on her belief about Y, that is, $P(Y = H \mid X)$

- How to compute $P(Y = H \mid X)$?
- Based on $P(\theta \mid X)$

$$
\begin{tikzcd}
P(\theta = 0.8 \mid X) \arrow[r, "0.8"] \arrow[dr] & Y = H \\
P(\theta = 0.2 \mid X) \arrow[r, "0.8"'] \arrow[ur] & Y = T
\end{tikzcd}
$$

- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

Flips the coin, sees a head $X = H$

- $\Pr(Y = H \mid X = H)$**?**

Flips the coin, sees a tail $X = T$

- $\Pr(Y = H \mid X = T)$**?**

Bayesian modeling

Warm up

Beliefs about the weight θ

$P(\theta \mid X)$	$\theta = 0.8$	$\theta = 0.2$
H	0.97	0.03
T	0.69	0.31

Bayesian modeling
Warm up

Beliefs about the test coin flip Y

$P(Y\|X)$	**Y = H**	**Y = T**
H	0.78	0.22
T	0.62	0.38

Flips the coin, sees a head $X = H$

- Report $X' =$ **?**

Flips the coin, sees a tail $X = T$

- Report $X' =$ **?**

Bayesian modeling

Warm up

Beliefs about the test coin flip Y

$P(Y \mid X)$	**Y = H**	**Y = T**
H	0.78	0.22
T	0.62	0.38

- Reward $R(X, Y) = 1$ if $X = Y$

  $R(X, Y) = 0$ if $X \neq Y$

Always report $X' = H$

Bayesian modeling

Warm up

Beliefs about the test coin flip Y

$P(Y \mid X)$	**Y = H**	**Y = T**
H	0.78	0.22
T	0.62	0.38

Reward

- $R(X, Y) = 1$ if $X = Y = H$
- $R(X, Y) = 10000$ if $X = Y = T$
- $R(X, Y) = 0$ if $X \neq Y$

Bayesian modeling

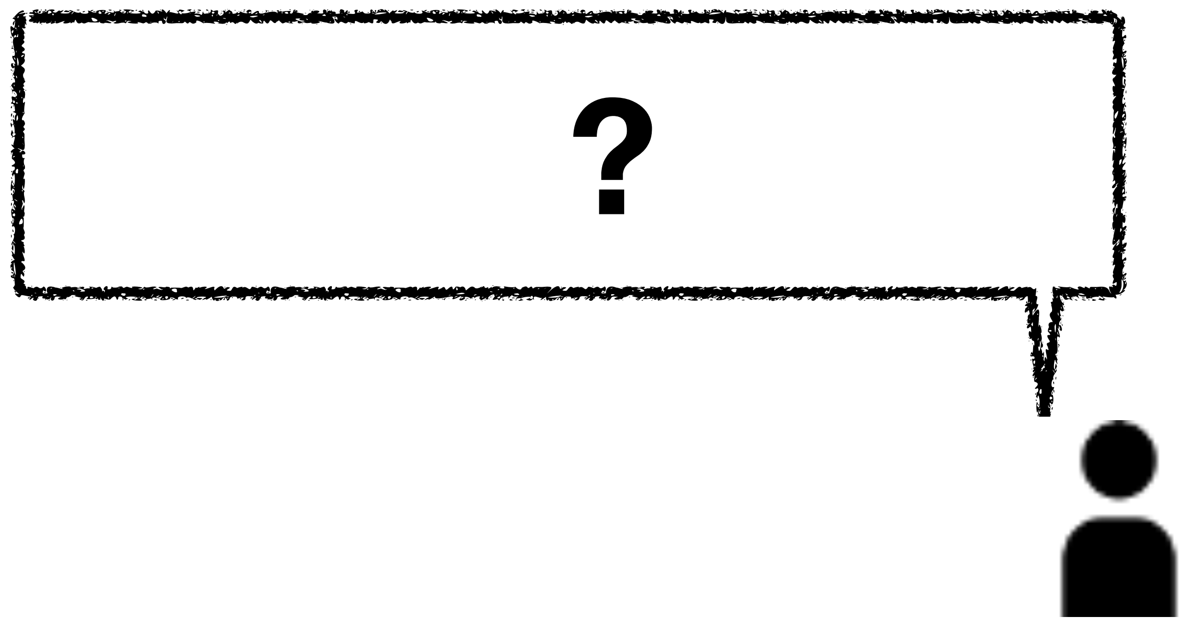
Warm up

Beliefs about the test coin flip Y

$P(Y\|X)$	**Y = H**	**Y = T**
H	0.78	0.22
T	0.62	0.38

Reward

- $R(X, Y) = 1$ if $X = Y = H$
- $R(X, Y) = 10000$ if $X = Y = T$
- $R(X, Y) = 0$ if $X \neq Y$



Bayesian modeling

Warm up

Beliefs about the test coin flip Y

$P(Y \mid X)$	**Y = H**	**Y = T**
H	0.78	0.22
T	0.62	0.38

Reward

- $R(X, Y) = 1$ if $X = Y = H$
- $R(X, Y) = 10000$ if $X = Y = T$
- $R(X, Y) = 0$ if $X \neq Y$

Always report $X' = T$

Bayesian modeling

Warm up

Beliefs about the test coin flip Y

$P(Y \mid X)$	**Y = H**	**Y = T**
H	0.78	0.22
T	0.62	0.38

Goal: design reward $R(X, Y)$

s.t.

Report $X' = H$ when seeing $X = H$

Report $X' = T$ when seeing $X = T$

Logarithmic scoring rule

- Reward $R(X, Y) = \log(P(Y | X))$
- Always give the true coin flip result

R(X,Y)	Y = H	Y = T
H	$\log 0.78$	$\log 0.22$
T	$\log 0.62$	$\log 0.38$

Logarithmic scoring rule

See $X = T$

- Expected reward of reporting T
  $= p_T \log p_T + (1 - p_T)\log(1 - p_T)$
- Expected reward of reporting H
  $= p_T \log p_H + (1 - p_T)\log(1 - p_H)$

R(X,Y)	Y = H	Y = T
H	$\log p_H$	$\log(1 - p_H)$
T	$\log p_T$	$\log(1 - p_T)$

Logarithmic scoring rule

See $X = T$

- Expected reward of reporting T
  $= p_T \log p_T + (1 - p_T)\log(1 - p_T)$
- Expected reward of reporting H
  $= p_T \log p_H + (1 - p_T)\log(1 - p_H)$
- Reporting $T-$ reporting H

  $= D_{KL}(p_L || p_H)$

R(X,Y)	Y = H	Y = T
H	$\log p_H$	$\log(1 - p_H)$
T	$\log p_T$	$\log(1 - p_T)$

Logarithmic scoring rule

See $X = T$

- Expected reward of reporting T
  $= p_T \log p_T + (1 - p_T) \log(1 - p_T)$
- Expected reward of reporting H
  $= p_T \log p_H + (1 - p_T) \log(1 - p_H)$
- Reporting $T -$ reporting H

  $= D_{KL}(p_L \| p_H) \geq 0$

R(X,Y)	Y = H	Y = T
H	$\log p_H$	$\log(1 - p_H)$
T	$\log p_T$	$\log(1 - p_T)$

Lemma: $D_{KL}(p \| q) \geq 0$

Bayesian modeling

Summary

Key idea:

The **loss** in the reward when manipulating data = KL divergence

reward(reporting D) - reward(reporting $f(D) = D'$) = KL divergence

- Can be extended to general Bayesian machine learning problems.

Data valuation by the log scoring rule

Reward a dataset D using a test dataset T

- Use logarithmic scoring rule
  $R(D, T) = \log(P(T|D))$
- Observe D and report D', the loss in the expected reward
  $= D_{KL}(P(T|D) \parallel P(T|D')) \geq 0$

R	T_1	T_2	$\cdots$	
D_1				
D_2		$\log P(T\|D)$		
			$\cdots$	
$\cdots$				

Data valuation by the log scoring rule

Reward a dataset D using a test dataset T

- Use logarithmic scoring rule
  $R(D, T) = \log(P(T|D))$
- Observe D and report D', the loss in the expected reward
  $= D_{KL}(P(T|D) \parallel P(T|D')) \geq 0$

R	T_1	T_2	$\cdots$	
D_1				
D_2		$\log P(T\|D)$		
			$\cdots$	
$\cdots$				

Theorem: By using log scoring rule $R(D, T) = \log(P(T|D))$, we have

$\mathbf{E}_T\left[R(D', T) \,|\, D\right] \leq \mathbf{E}_T\left[R(D, T) \,|\, D\right]$, for any **possible** D'

Outline

A data valuation method that prevents data manipulations

- Bayesian modeling & the log scoring rule
- Computing the log scoring rule for Bayesian machine learning
- Sensitivity analysis
- Summary & extension

Bayesian machine learning

- A ML model with parameter θ
- A probability distribution of θ
- $\theta \sim P(\theta)$, update $P(\theta \mid D) \propto P(\theta)P(D \mid \theta)$
- Generate predictions using $P(\theta \mid D)$
  - Maximum A Posteriori (MAP) estimation, $\theta^* = \arg\max_{\theta} P(\theta \mid D)$

<u>Assumption:</u> For any dataset D, the posterior $P(\theta \mid D)$ is computable

Data valuation for Bayesian ML

Suppose a data provider collects data $D = \{x_i\}_{i=1}^{n}$ with $x_i \sim P(x \mid \theta)$

- We have a test dataset $T = \{x_j\}_{j=1}^{m}$ with $x_j \sim P(x \mid \theta)$ drawn **independently**

Goal: design a valuation function $R(D, T)$ such that

$\mathbf{E}_{\theta,T}\left[R(f(D), T) \mid D\right] \leq \mathbf{E}_{\theta,T}\left[R(D, T) \mid D\right]$, for any manipulation $f(\,\cdot\,)$

Theorem: We can use the log scoring rule $R(D, T) = \log(P(T \mid D))$.

- How do we compute $P(T \mid D)$?

Computing the log scoring rule

How do we compute $P(T \mid D)$?

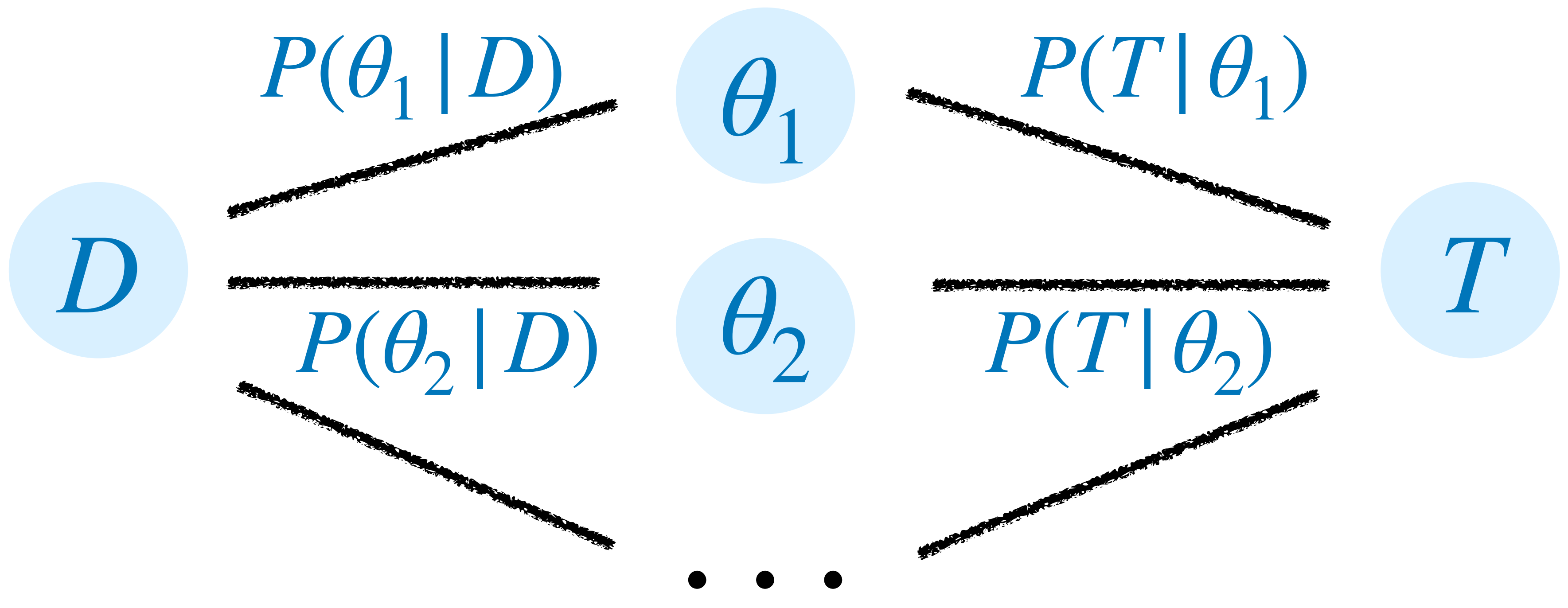
- The simplest approach: generate predictive distribution using the posterior $P(\theta \mid D)$

Computing the log scoring rule

How do we compute $P(T \mid D)$?

- The simplest approach: generate predictive distribution using the posterior $P(\theta \mid D)$

Lemma: $P(T \mid D) = \int_{\theta} P(T \mid \theta) P(\theta \mid D) \, d\theta$.

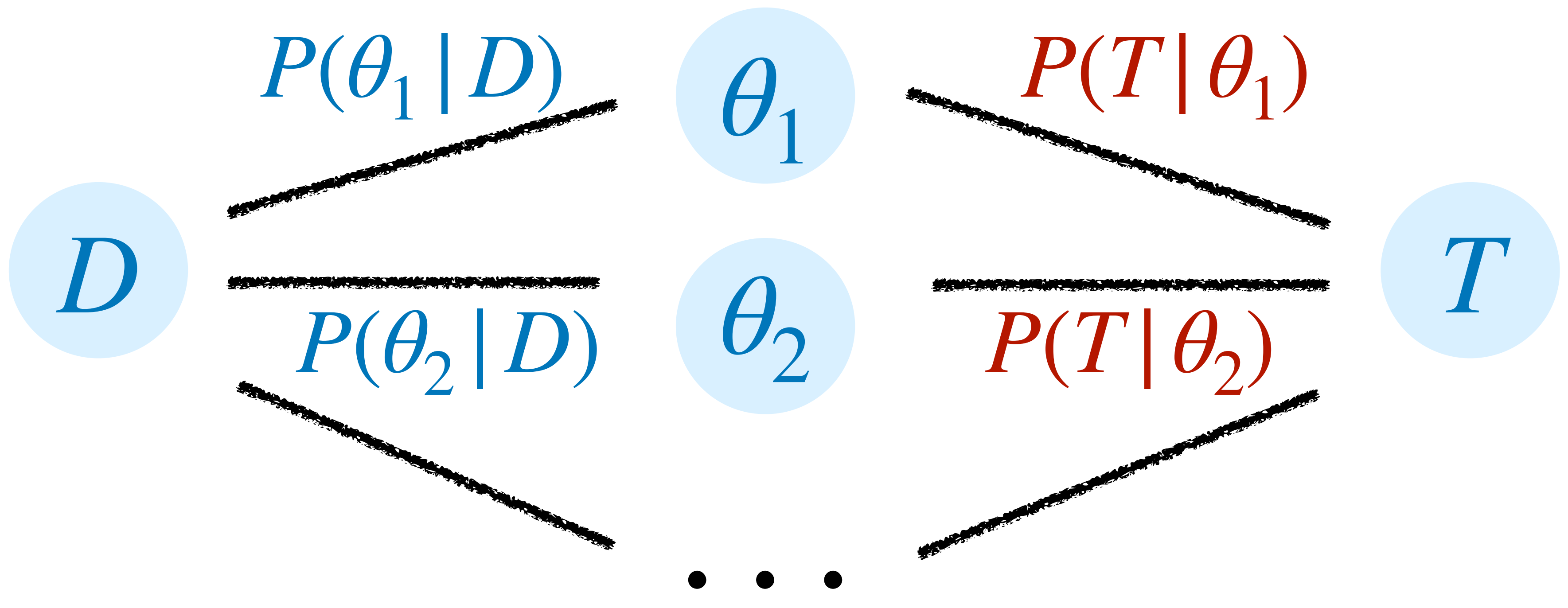


Computing the log scoring rule

How do we compute $P(T \mid D)$?

- The simplest approach: generate predictive distribution using the posterior $P(\theta \mid D)$

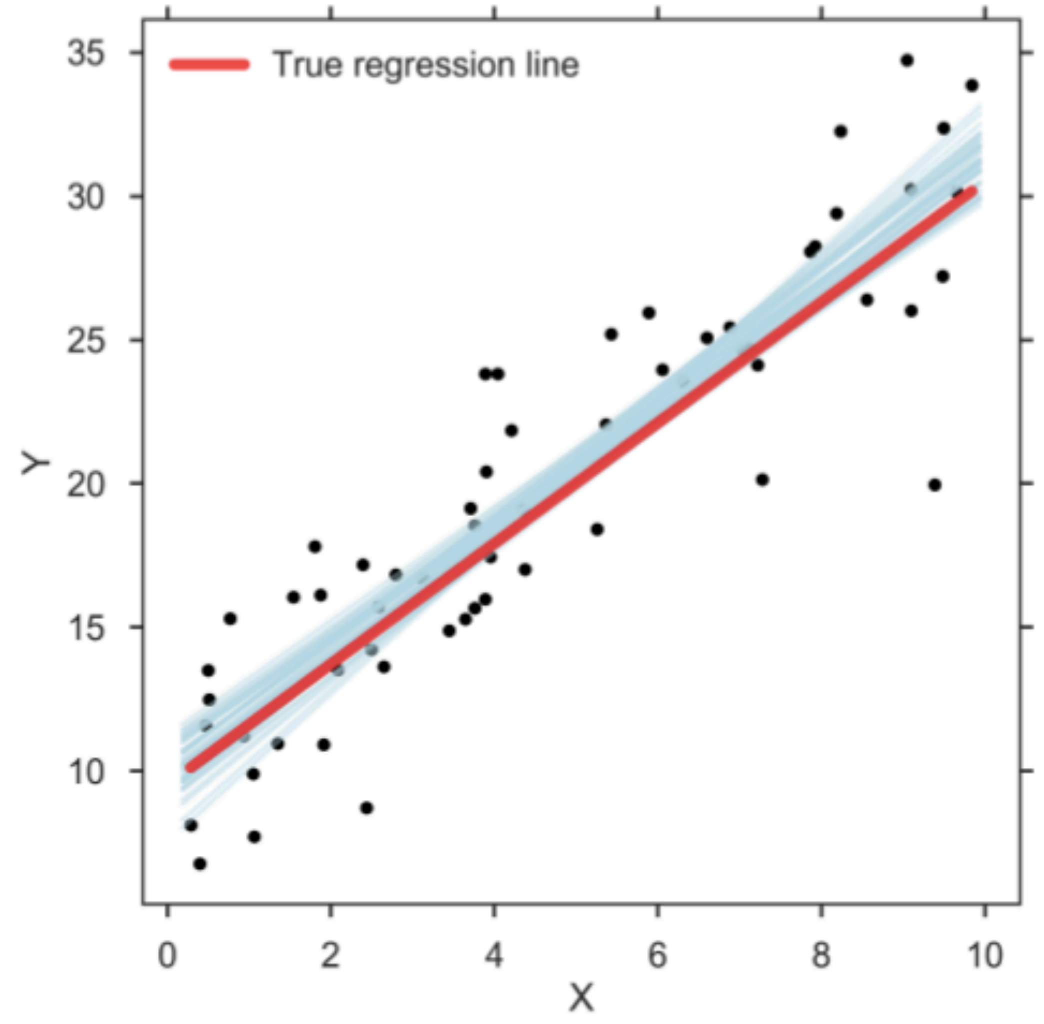
Lemma: $P(T \mid D) = \int_{\theta} P(T \mid \theta) P(\theta \mid D)\, d\theta.$



Problem: need to have a model for $P(T \mid \theta)$

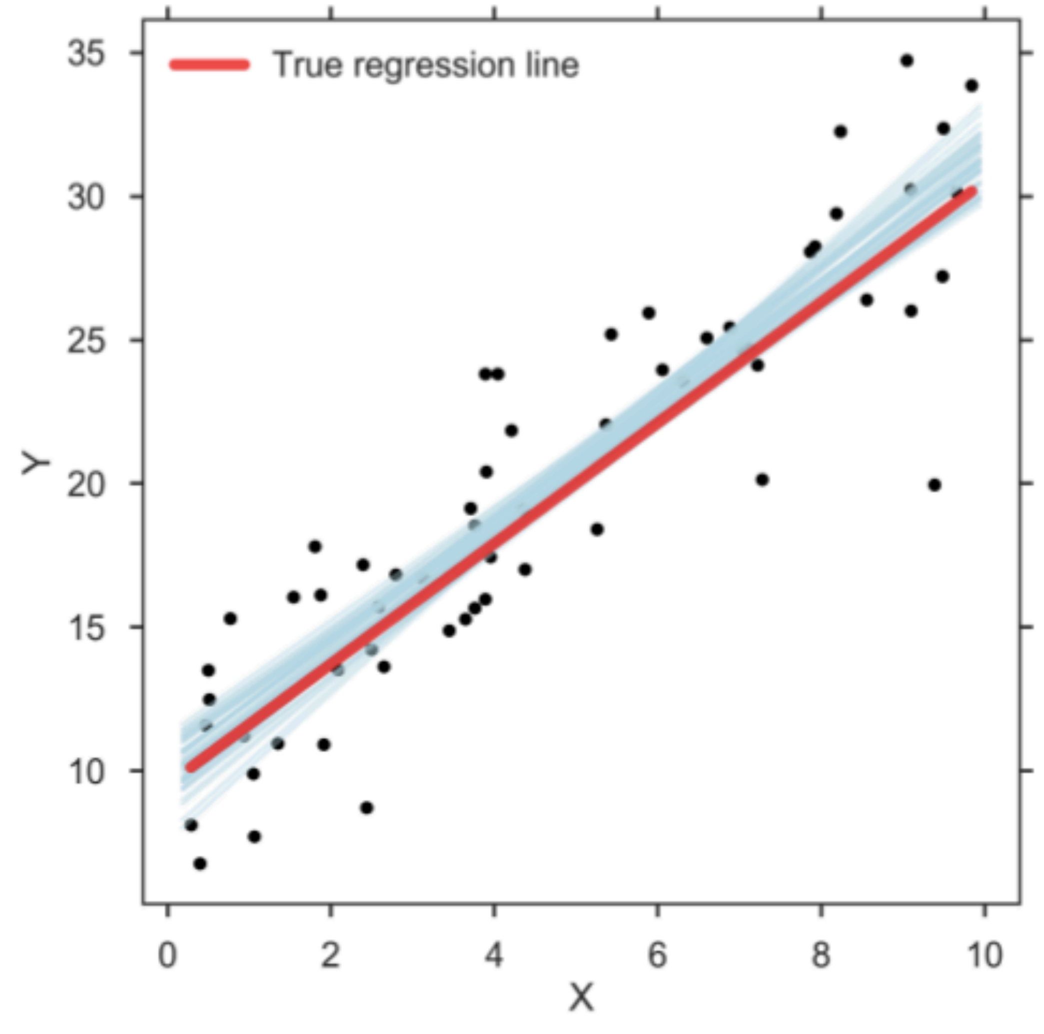
Computing the log scoring rule

- **Problem:** for some Bayesian ML problem, $P(T \mid \theta)$ **not fully modeled**
- Consider Bayesian linear regression: data point $(\mathbf{x}_i, y_i)$
- $y_i = \theta^T \mathbf{x}_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$
- Prior $\theta \sim N(\mu_0, \sigma_0^2)$, can compute posterior $P(\theta \mid D)$ in closed form



Computing the log scoring rule

- **Problem:** for some Bayesian ML problem, $P(T \mid \theta)$ **not fully modeled**
- Consider Bayesian linear regression: data point $(\mathbf{x}_i, y_i)$
- $y_i = \theta^T \mathbf{x}_i + \varepsilon_i$ with $\varepsilon_i \sim N(0,\ \sigma_\varepsilon^2)$
- Prior $\theta \sim N(\mu_0,\ \sigma_0^2)$, can compute posterior $P(\theta \mid D)$ in closed form



Distribution of $\mathbf{x}_i$ not specified

Computing the log scoring rule

Question: Can we still use $R(D, T) = \log(P(T \mid D))$ when the data distribution $P(T \mid \theta)$ **is not fully specified**?

- Yes! But a variant of the log scoring rule

- Don't need $P(T \mid \theta)$
- Only need $P(\theta \mid D)$ and $P(\theta \mid T)$

Computing the log scoring rule

- Reward $R(D, T) = \log P(T \mid D) - \log P(T)$

$$\equiv \log P(T \mid D) - \text{a constant}$$

Computing the log scoring rule

- Reward $R(D, T) = \log P(T \mid D) - \log P(T)$

  $\equiv \log P(T \mid D) -$ a constant

Does not depend on D

Theorem: By using $R(D, T) = \log P(T \mid D) - \log P(T)$, we have

$\mathbf{E}_{\theta,T}\,[R(D', T) \mid D] \leq \mathbf{E}_{\theta,T}\,[R(D, T) \mid D]$, for any **possible** D'

Computing the log scoring rule

- Reward $R(D, T) = \log P(T \mid D) - \log P(T)$

$$= \log \left(P(T \mid D) / P(T) \right)$$

$\equiv \log P(T \mid D) -$ a constant

Computing the log scoring rule

- Reward $R(D,T) = \log P(T|D) - \log P(T)$

$$= \log \left(P(T|D)/P(T)\right)$$

$\equiv \log P(T|D) -$ a constant

Lemma (Kong and Schoenebeck, 2018): When the data points in D and T are drawn independently from $P(x|\theta)$,

$$\frac{P(T|D)}{P(T)} = \int_\theta \frac{P(\theta|T)P(\theta|D)}{P(\theta)}\, d\theta.$$

Proof: $\displaystyle \frac{P(T|D)}{P(T)} = \frac{\int_\theta P(T|\theta)P(\theta|D)\, d\theta}{P(T)} = \int_\theta \frac{P(T|\theta)}{P(T)} \cdot P(\theta|D)\, d\theta = \int_\theta \frac{P(\theta|T)}{P(\theta)} \cdot P(\theta|D)\, d\theta$

Computing the log scoring rule

- Reward $R(D,T) = \log P(T|D) - \log P(T)$

$$= \log \left(P(T|D)/P(T)\right)$$

$\equiv \log P(T|D) -$ a constant

Lemma (Kong and Schoenebeck, 2018): When the data points in D and T are drawn independently from $P(x|\theta)$,

$$\frac{P(T|D)}{P(T)} = \int_\theta \frac{P(\theta|T)P(\theta|D)}{P(\theta)}\, d\theta.$$

Bayes' Lemma

Proof: $\frac{P(T|D)}{P(T)} = \frac{\int_\theta P(T|\theta)P(\theta|D)\, d\theta}{P(T)} = \int_\theta \frac{P(T|\theta)}{P(T)} \cdot P(\theta|D)\, d\theta = \int_\theta \frac{P(\theta|T)}{P(\theta)} \cdot P(\theta|D)\, d\theta$

Computing the log scoring rule

- Reward $R(D,T) = \log P(T|D) - \log P(T)$

$\equiv \log P(T|D) -$ a constant

$$= \log \left(P(T|D)/P(T)\right)$$

$$= \log \int_\theta \frac{P(\theta|T)P(\theta|D)}{P(\theta)}\, d\theta$$

- Only needs the prior and the posteriors
- Easy to compute for a class of widely-used distributions: **exponential families**
    - Bernoulli, Gaussian, Multinomial, Dirichlet, Gamma, Poisson, Beta

Computing the log scoring rule

Exponential family

- Easy to compute for a class of widely-used distributions: **exponential families**
  - Bernoulli, Gaussian, Multinomial, Dirichlet, Gamma, Poisson, Beta

Lemma [Chen et al. 2020]: The reward $R(D, T) = \log \int_\theta \frac{P(\theta | T) P(\theta | D)}{p(\theta)} \, d\theta$ can be computed in O(# of data points in D and T) time if the data generating distribution $P(x | \theta)$ is in an **exponential family** and $P(\theta)$ is a **conjugate prior** for $P(x | \theta)$.

Computing the log scoring rule

Exponential family

- For the coin flips example, we have $x \sim \text{Ber}(x \mid \theta)$ in an exponential family
- Suppose $P(\theta) = \text{Beta}(a, b)$
- Only need to count the **# of heads** in the datasets: T has a_T heads and b_T tails, D has a_D heads and b_D tails, then

$$R(D, T) = \frac{B(a + a_T, b + b_T)B(a + a_D, b + b_D)}{B(a, b)B(a + a_T + a_D, b + b_T + b_D)}$$

where $B(\cdot\,,\cdot)$ is the Beta function.

Outline

A data valuation method that prevents data manipulations

- Bayesian modeling & the log scoring rule
- Computing the log scoring rule for Bayesian machine learning
- Sensitivity analysis
- Summary & extension

Sensitivity analysis

Theorem: By using log scoring rule $R(D, T) = \log(P(T \mid D))$, **we have**

$\mathbf{E}_{\theta,T}\left[R(D', T) \mid D\right] \leq \mathbf{E}_{\theta,T}\left[R(D, T) \mid D\right]$, for any **possible** D'

- Only guarantee weak inequality (can be achieved by a constant payment)
- Strictly better?

Sensitivity analysis

(Chen et al. 2020) sensitivity analysis

- Undesirable manipulation: D' that gives a different posterior distribution

$$P(\theta \mid D') \neq P(\theta \mid D)$$

- Discrete distribution: manipulation is strictly worse if the test dataset T has enough correlation with D

For discrete $P(x \mid \theta)$ and $P(\theta)$, assuming that different θ lead to different data distributions, any undesirable manipulation is strictly worse if the number of the test data points $|T| \geq |\Theta| - 1$

Sensitivity analysis

(Chen et al. 2020) sensitivity analysis

- Continuous distribution: depend on the model, can detect certain manipulations
- E.g., estimate the mean of a Gaussian distribution: $x \sim N(\theta,\ 1)$
- Can detect the change in the #data points (duplicating data, withholding data)
- But not the change in the values of the data points

Outline

A data valuation method that prevents data manipulations

- Bayesian modeling & the log scoring rule
- Computing the log scoring rule for Bayesian machine learning
- Sensitivity analysis
- Summary & extension

Summary

1. A data valuation method based on the log scoring rule

- prevents data manipulation

2. Easy to compute for a large class of BML problems

Pros and cons

A data valuation method based on the log scoring rule

- Pros: strong theoretical guarantee

Theorem: By using log scoring rule $R(D, T) = \log(P(T \mid D))$, **we have**

$\mathbf{E}_{\theta,T}\left[R(D', T) \mid D\right] \leq \mathbf{E}_{\theta,T}\left[R(D, T) \mid D\right]$, for any **possible** D'

- Cons:
  - Randomized, truthful in expectation
  - Unbounded $R(D, T) = \log(P(T \mid D))$

References

Scoring rules:

Gneiting and Raftery, “Strictly Proper Scoring Rules, Prediction, and Estimation”

Compute by posteriors:

Kong and Schoenebeck, “Water from Two Rocks: Maximizing the Mutual Information”, EC 2018

Exponential family & sensitivity analysis:

Chen, Shen, and Zheng, “Truthful Data Acquisition via Peer Prediction”, Neurips 2020

Thanks & Questions?

Shuran: collect truthful data from strategic/self-interested agents

NEXT

James: ML-as-service market and competition among ML vendors

Shuran: collect truthful da
interested

15 mins break!

NEXT

James: ML-as-service market and competition among ML vendors