ILLINOIS INSTITUTE OF TECHNOLOGY

Privacy Preserving Bag Preparation for Learning from Label Proportion

11.6.2018

CS-591 Thesis Defense

Xinzhou Yan

M.S CDS & O.R

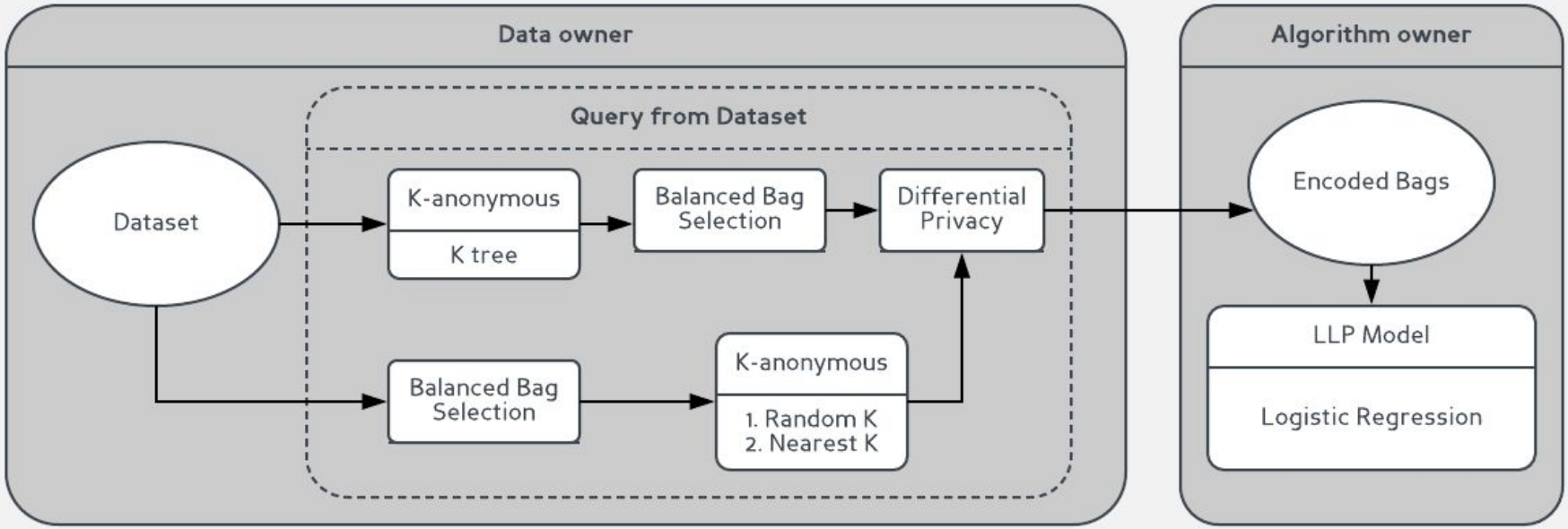
Main Target

- **Which Machine Learning model to choose?**
  (LLP: Learning from label proportion framework)
- **How to apply PPDM to LLP model?**
  (PPDM: Privacy preserving Data Mining standards)
- **How to ensure accuracy loss for PPDM LLP model?**

PPDM LLP Model

PPDM: Privacy preserving Data Mining standards
LLP: Learning from label proportion framework

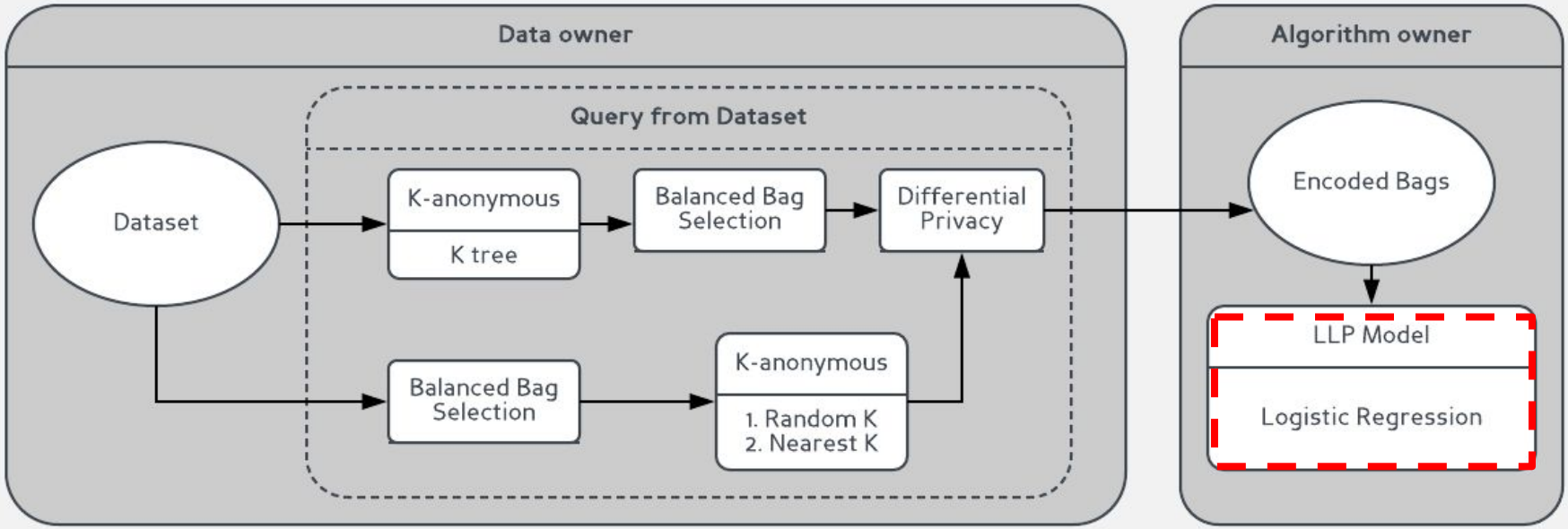
- No sensitive information can be retrieved from the dataset
- No record can be traced traced in the dataset.



LLP Model

PPDM: Privacy preserving Data Mining standards
LLP: Learning from label proportion framework

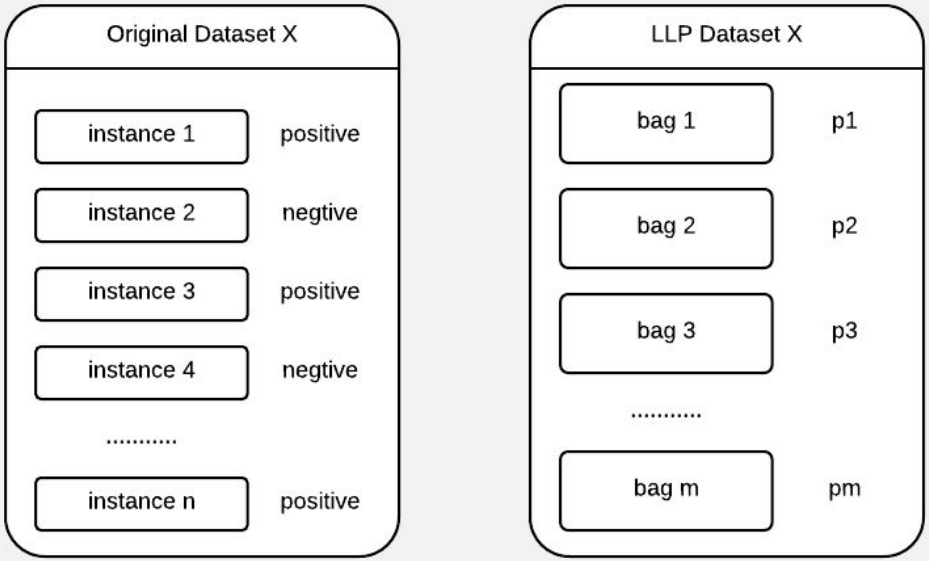
- No sensitive information can be retrieved from the dataset
- No record can be traced traced in the dataset.



LLP model

LLP is a Learning from label proportion framework which enable the learner to train on unlabeled data

- In LR setting, each record has to be labeled.
- In LLp setting, Instances are put into bags with prior label distribution



LLP model

Use Logistic regression- LLP model to fit the dataset
Each bag's probability $p(X_i, p_i|\theta)$ **can be modeled as Beta distribution**

- X represents the feature matrix, Xi represents the feature matrix of ith bag
- pi represents the ith prior label distribution of ith bag
- p(Xij, θ) is the logistic function, which has domain [0, 1]
- p(Xi, θ) is the average label distribution which has domain [0, 1]

$$p(X_i, p_i|\theta) = p(X_i;\theta)^{p_i}(1-p(X_i;\theta))^{1-p_i}$$

$$p(X, p|\theta) = \prod_{i=1}^{n} p(X_i;\theta)^{p_i}(1-p(X_i;\theta))^{1-p_i}$$

$$p(\theta|X, p) \propto p(\theta)p(X, p|\theta) \propto p(X, p|\theta)$$

LLP model

Use Logarithmic Likelihood function to estimate the parameter
The optimal parameter is also the optimal parameter for the loss function

- X represents the feature matrix, Xi represents the feature matrix of ith bag
- Add up regularization term to ensure θ is close to zero vector.

$$l(\theta) = log(p(X, p|\theta))$$

$$= log \prod_{i=1}^{n} p(X_i; \theta)^{p_i} (1 - p(X_i; \theta))^{1-p_i}$$

$$= \sum_{i=1}^{n} (p_i log(p(X_i; \theta))) + \sum_{i=1}^{n} ((1 - p_i) log(1 - p(X_i; \theta)))$$

$$loss(P, p(X, \theta)) = -l(\theta) + \frac{\lambda \theta^2}{2}$$

LLP model

Prediction of a single instance in LLP model:

$$p(X_{i,j};\theta) = \frac{1}{1+e^{\theta^T X_{i,j}}}$$

Prediction of the average label distribution in the ith bag of the LLP model:

$$p(X_i;\theta) = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{1+e^{\theta^T X_{i,j}}}$$

Loss function:

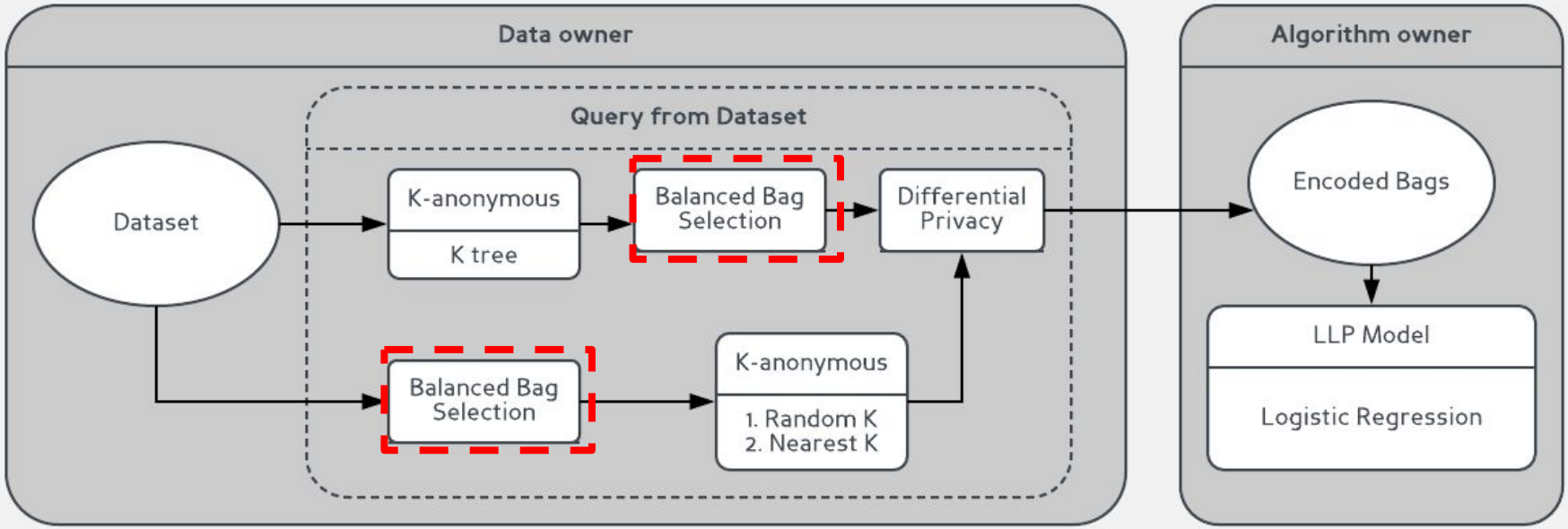
$$loss(P, p(X,\theta)) = -l(\theta) + \frac{\lambda\theta^2}{2}$$

$$\frac{\partial loss(P, p(X,\theta))}{\partial\theta} = \frac{1}{m}\sum_{i=1}^{n}\left(\frac{p_i - p(X_i;\theta)}{p(X_i;\theta)(1-p(X_i;\theta))}\right)\sum_{j=1}^{m} p(X_{i,j};\theta)(1-p(X_{i,j};\theta))X_{i,j} + \lambda\theta$$

Balanced Bag Selection

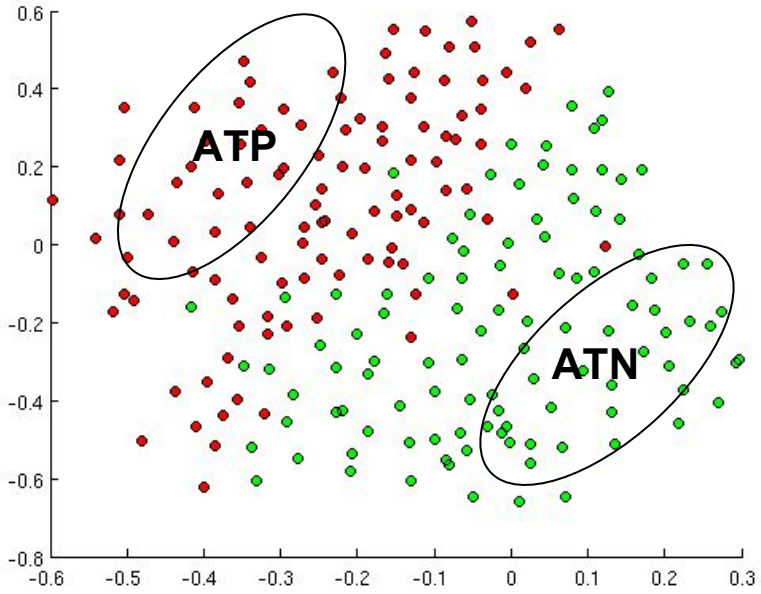
PPDM: Privacy preserving Data Mining standards
LLP: Learning from label proportion framework

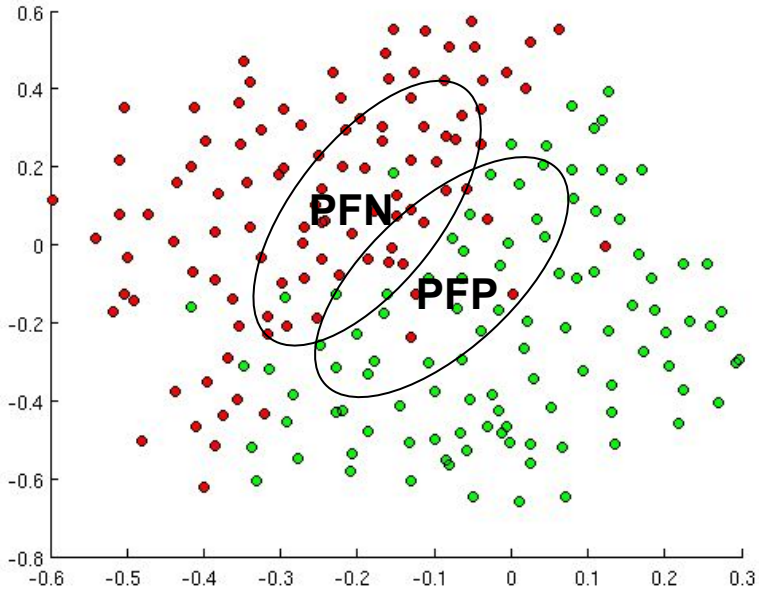
- No sensitive information can be retrieved from the dataset
- No record can be traced traced in the dataset.



Balanced Bag Selection

We define highly confident instances as ”absolute true positive”(ATP) or ”absolute true negative”(ATN), some lowly confident instances as ”possible false positive”(PFP) or”possible false negative”(PFN).





[1] Image downloaded from Andrew Ng’s website

Experiment

Select bags i.i.d from confident , unconfident instances. The loss function punishes both the misclassification and low-precision model .

- Misclassification is punished by the wide margin, ATP&ATN
- The low precision is punished by the narrow margin: PFN&PFP

Table 4.3. Accuracy/F1 of different models

model	Train on	Positive F1	Negative F1	Average F1
logistic regression	Adult	72	91	86
LLP	Adult	45	85	75
Balanced-LLP	Adult	57	89	81
logistic regression	Instagram	64	96	92
LLP	Instagram	27	86	84
Balanced-LLP	Instagram	61	95	91

Experiment

Table 4.1. Average F1 of Bag selection Methods

Bag selection	Train on	Average F1
I.I.D Selection	Adult	0.75
ATP/ATN PFP/PFN	Adult	0.81
ATP/PFP PFN/ATN	Adult	0.81
ATP/ATN	Adult	77
PFP/PFN	Adult	72
I.I.D Selection	Instagram	0.86
ATP/ATN PFP/PFN	Instagram	0.91
ATP/PFP PFN/ATN	Instagram	0.87
ATP/ATN	Instagram	0.88
PFP/PFN	Instagram	0.30

PPDM Standards

PPDM: privacy preserving data mining.
PPDM ensures necessary security when querying the database.

- K-anonymous: any record in the database cannot be distinguished from other k-1 records.
- L-DIversity: at least L distinct labels for the records in each block
- Differential privacy: add Laplace noise to the dataset to ensure:

$$e^{-\varepsilon} \leq \frac{pr(M(X')=C)}{pr(M(X)=C)} \leq e^{\varepsilon}$$

X' is neighbour dataset , M is the query function, and C is the query result

PPDM Standards

Enough sensitivity, enough security for the differential privacy model

- Sensitivity is the maximal difference of the query result between two neighbour datasets.
- Add laplace noise to the query result to ensure differential privacy.

$$let\ \ \Delta f = max|M(X) - M(X')|,\ \ z \sim Lap(b),\ \ b = \Delta f/\varepsilon$$

$$\frac{pr(M(X')=C)}{pr(M(X)=C)} = \frac{\frac{1}{2b} e^{-|M(X')|/b}}{\frac{1}{2b} e^{-|M(X)|/b}} = e^{(|M(X)| - |M(X')|)/b} \le e^{(\Delta f)/b} \le e^{\varepsilon}$$

$$reversely,\ \ \frac{pr(M(X')=C)}{pr(M(X)=C)} \ge e^{(-\Delta f)/b} \ge e^{-\varepsilon}$$

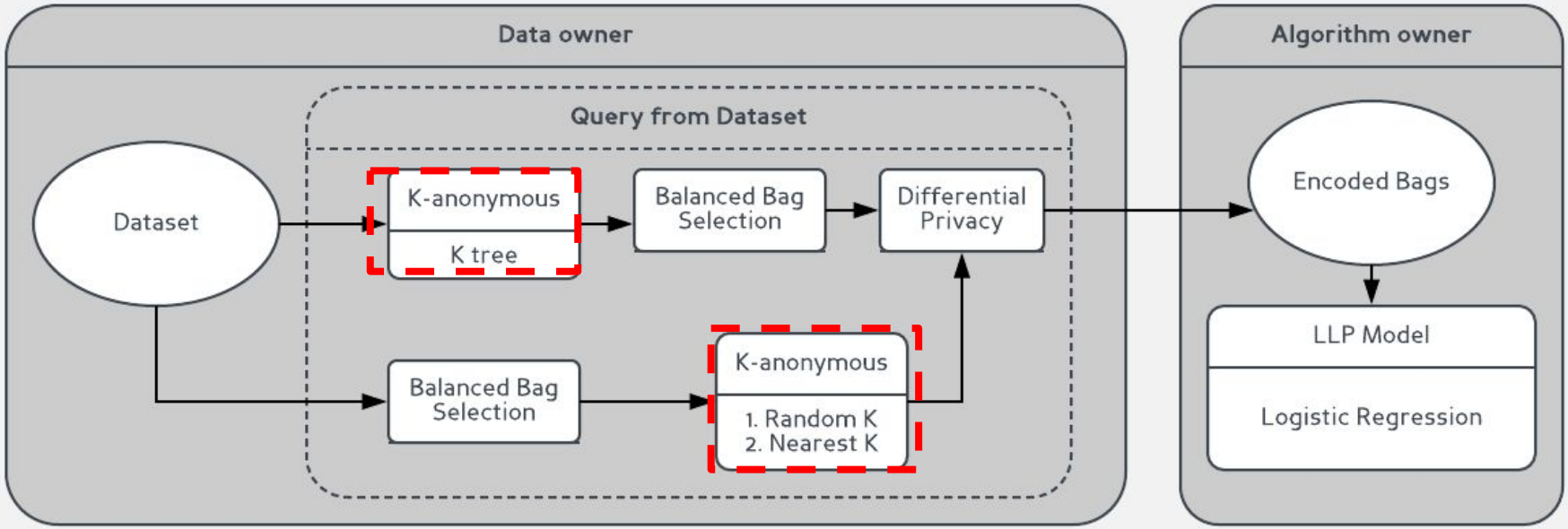
$$e^{-\varepsilon} \le \frac{pr(M(X')=C)}{pr(M(X)=C)} \le e^{\varepsilon}$$

[2] Deduction is quoted from Dwork's work.

K-Anonymous

Aggregate k instances in each bag and get mean feature vectors.

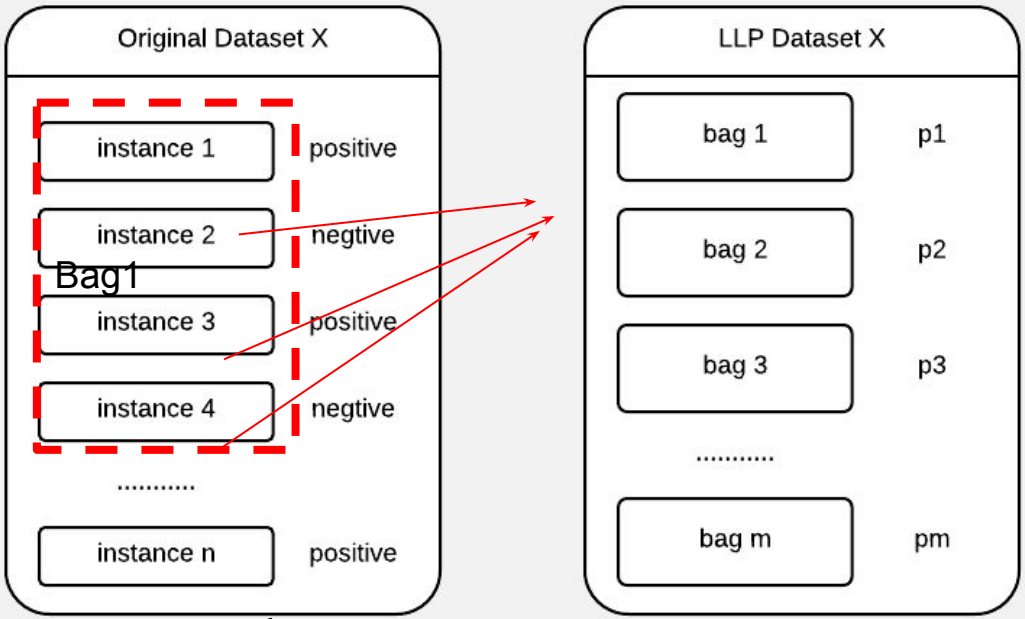
- Replace original instance with mean feature vector
- Label distribution unchanged



K-anonymous

Aggregation both positive and negative instances together

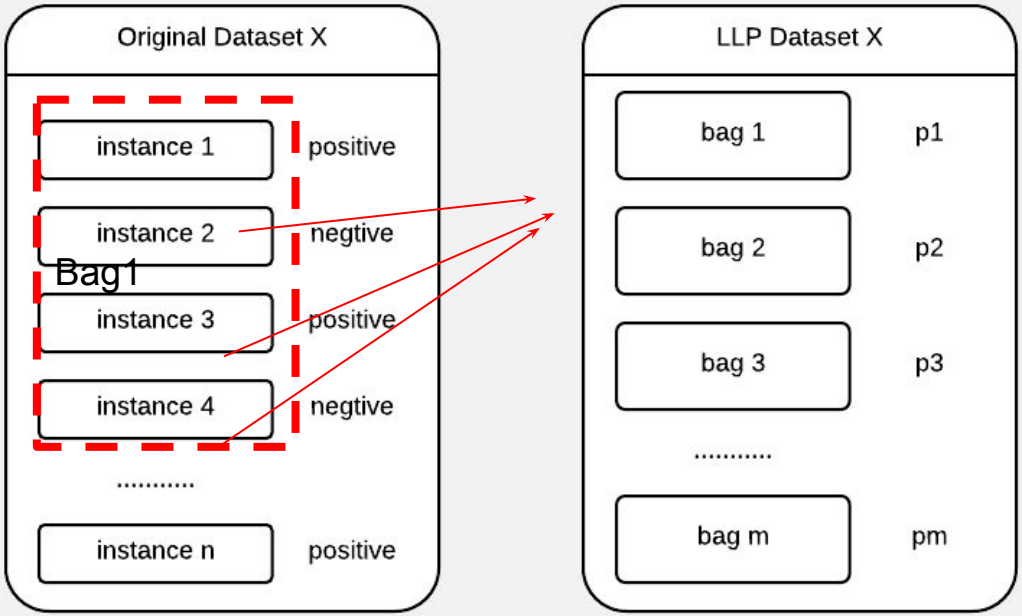
- Random K aggregation: randomly aggregate every k instances in each bag.
- Nearest K aggregation: find every k instances closest to one among k in each bag and aggregate them.



K-anonymous

Aggregation both positive and negative instances together

- Random K aggregation: randomly aggregate every k instances in each bag.
- Nearest K aggregation: find every k instances closest to one among k in each bag and aggregate them.



Accuracy Loss for LR

LR model guarantees accuracy loss when aggregation +/- instances separately.
LLP model guarantees accuracy loss and aggregate +/- instances together.

$$
\begin{aligned}
f(\theta) &= loss(\theta) \\
&= -\sum_{i=1}^{n} (y_i log P(x_i, \theta) + (1 - y_i) log(1 - P(x_i, \theta))) \\
&= -\sum_{i=1}^{n} (y_i log \frac{1}{1 + e^{\theta^T x_i}} + (1 - y_i) log(1 - \frac{1}{1 + e^{\theta^T x_i}})) + \frac{1}{2}\lambda\theta^2 \\
f'(\theta) &= -\sum_{i=1}^{n} (-y_i \frac{e^{\theta^T x_i}}{1 + e^{\theta^T x_i}} x_i + (1 - y_i) \frac{1}{1 + e^{\theta^T x_i}} x_i) + \lambda\theta \\
&= -\sum_{n=i}^{n} \frac{1}{1 + e^{\theta^T x_i}} x_i - y_i x_i + \lambda\theta \\
f(0) &= -\sum_{i=1}^{n} (y_i(-1) + (1 - y_i)(-1)) = n \\
f'(0) &= -\sum_{i=1}^{n} (\frac{1}{2} x_i - y_i x_i) = -\sum_{i=1}^{n} (\frac{1}{2} - y_i) x_i \\
f(\theta) &\approx f(0) + \theta^T f'(0) = n - \theta^T \sum_{i=1}^{n} (\frac{1}{2} - y_i) x_i \\
&= n + \frac{1}{2}\theta^T \sum_{i=1,pos}^{n} x_i = n + \frac{1}{2}\theta^T \sum_{i=1,pos}^{n} \bar{x}
\end{aligned}
$$

Accuracy Loss for LLP

LR model guarantees accuracy loss when aggregation +/- instances separately.

LLP model guarantees accuracy loss and aggregate +/- instances together.

$$
\begin{aligned}
f(\theta) &= loss(\theta) \\
&= -\sum_{i=1}^{n}(p_i log(p(X_i;\theta))) + \sum_{i=1}^{n}((1-p_i)log(1-p(X_i;\theta))) \\
&\quad + \frac{1}{2}\lambda\theta^2 \\
f'(\theta) &= \frac{1}{m}\sum_{i=1}^{n}(\frac{P_i - p(X_i;\theta)}{p(X_i;\theta)(1-p(X_i;\theta))})\sum_{j=1}^{m}p(X_{i,j};\theta)(1-p(X_{i,j};\theta))X_{i,j} \\
&\quad + \lambda\theta \\
f(0) &= -\sum_{i=1}^{n}(p_i(-1) + (1-P_i)(-1)) = n \\
f'(0) &= \frac{1}{m}\sum_{i=1}^{n}(\frac{P_i - \frac{1}{2}}{\frac{1}{2}*\frac{1}{2}}\sum_{j=1}^{m}\frac{1}{2}*\frac{1}{2}X_{i,j}) \\
&= \frac{1}{m}\sum_{i=1}^{n}((P_i - \frac{1}{2})\sum_{j=1}^{m}X_{i,j}) \\
&= \sum_{i=1}^{n}((P_i - \frac{1}{2})\bar{X}_i) \\
f(\theta) &\approx f(0) + \theta^T f'(0) = n + \theta^T\sum_{i=1}^{n}((P_i - \frac{1}{2})\bar{X}_i)
\end{aligned}
$$

Data Introduction

Adult Dataset: 48842 records, 25%+, 75% -. Categorical and numerical features

- Discretizing data, dimension 338
- Balanced bag selection
- Threshold is 0.9, naive Bayes model
- 100 bags, 100 instances in each bag.
- Prior label distribution vector as label for bags

Instagram Hostile Comment: 30987 records, 13%+, 87%. Text categorical features.

- Discretizing data, dimension 6046
- Balanced bag selection
- Threshold is 0.9
- 100 bags, 100 instances in each bag.
- Prior label distribution vector as label for bags

eg:

'This is me, though, emotionally ;D', 'Good', ...,
'Well Yeah,But they dont have an actor or a new suit yet,sadly',
'Maybe Another MCU film before civil war',
"Robert Downey Jr. said there's a big announcement in a few days. Maybe it's about Spidey.

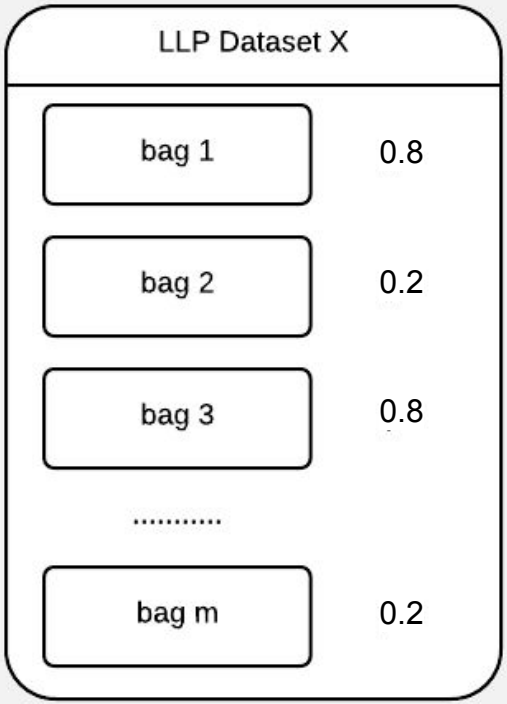
Data Introduction

Adult Dataset: 48842 records, 25%+, 75% -. Categorical and numerical features

- Discretizing data, dimension 338
- Balanced bag selection
- Threshold is 0.9, naive Bayes model
- 100 bags, 100 instances in each bag.
- Prior label distribution vector as label for bags

Instagram Hostile Comment: 30987 records, 13%+, 8

- Discretizing data, dimension 6046
- Balanced bag selection
- Threshold is 0.9
- 100 bags, 100 instances in each bag.
- Prior label distribution vector as label for bags

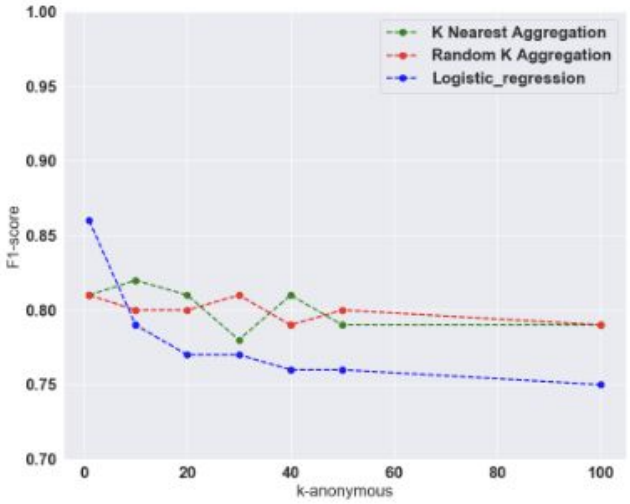


eg:

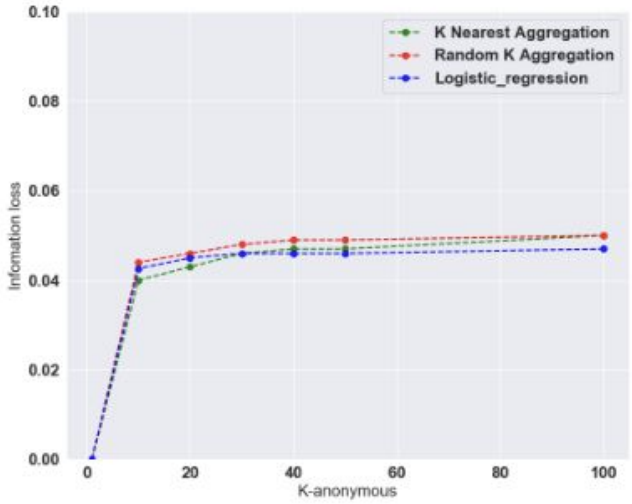
'This is me, though, emotionally ;D', 'Good', ...,
'Well Yeah,But they dont have an actor or a new suit yet,
'Maybe Another MCU film before civil war',
"Robert Downey Jr. said there's a big announcement in a few days. Maybe it's about Spidey.

Performance

LLP model is much more robust than LR model when L-diversity is satisfied.



(a) K-value V.S. F1-score

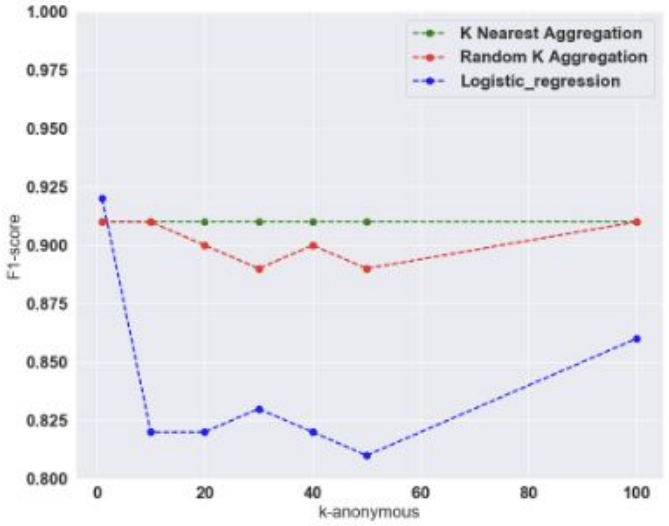


(b) Information loss V.S. K-value

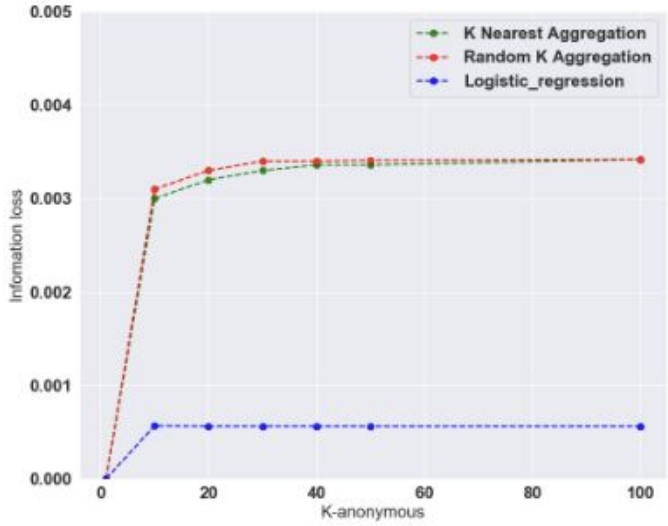
Figure 4.2. Adult Dataset

Performance

LR model loses accuracy loss at all.



(a) K-value V.S. F1-score



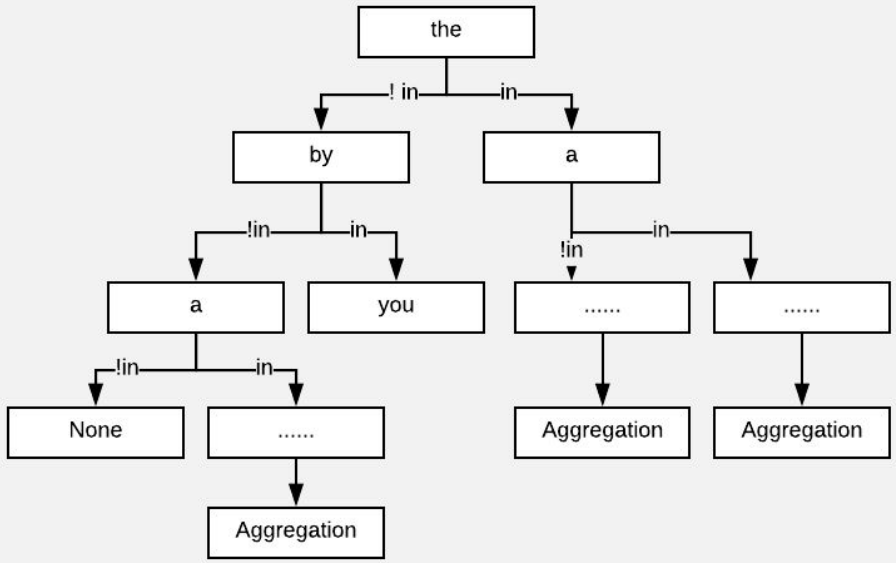
(b) Information loss V.S. K-value

Figure 4.3. Instagram Dataset

K-anonymous

K tree aggregation aggregates instances before putting into bags.
So the performance cannot be guaranteed by the formula above.

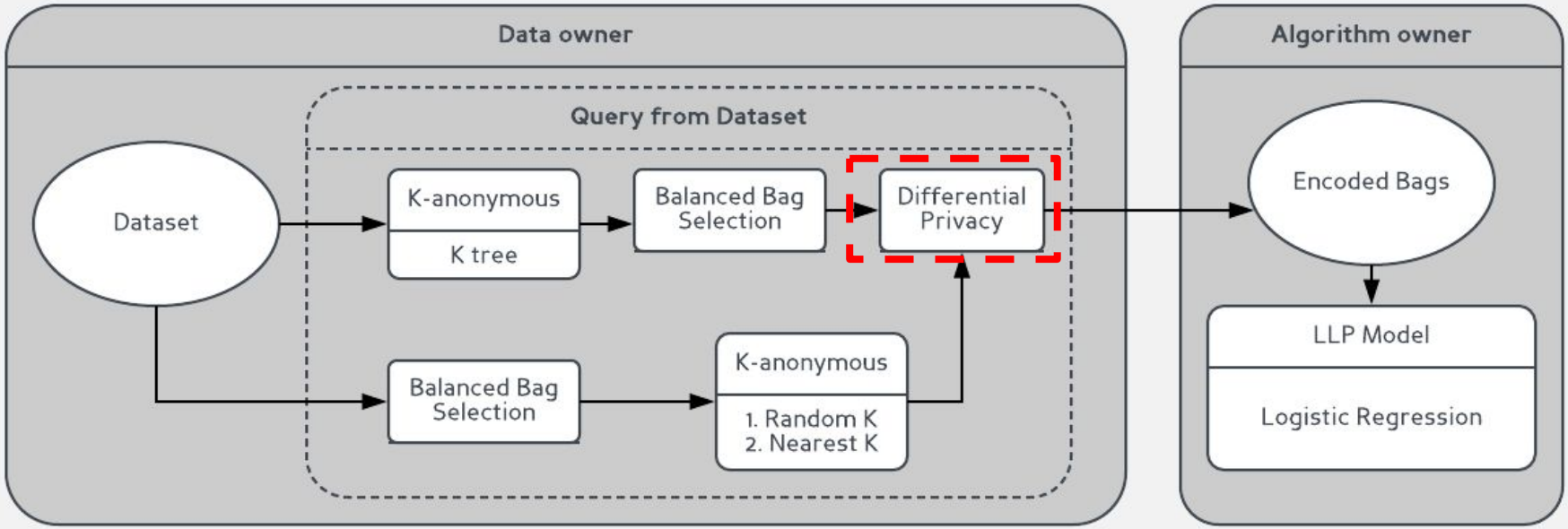
- K tree aggregates +/- instances separately, which cannot guarantee l-diversity.
- K tree allows big k value which equals to the maximal frequency of minority class' attributes.
- K tree is robust after applying differential privacy encoding.



Differential privacy

Ensure no record can be traced from the dataset

- Totally distorted feature matrix
- Totally distorted label distribution for each bag



Differential privacy

The sensitivity for the feature matrix:

- Query the dataset for the mean of k records' each attribute
- K is the anonymous value for the model
- The maximal difference of the query between 2 neighbour datasets is 1/k
- Prefer small ε and big k

The sensitivity for the prior label distribution:

- Query the dataset for each bag's label distribution
- m is the size of each bag
- The maximal difference of the query is 1/m
- Prefer small ε and big m

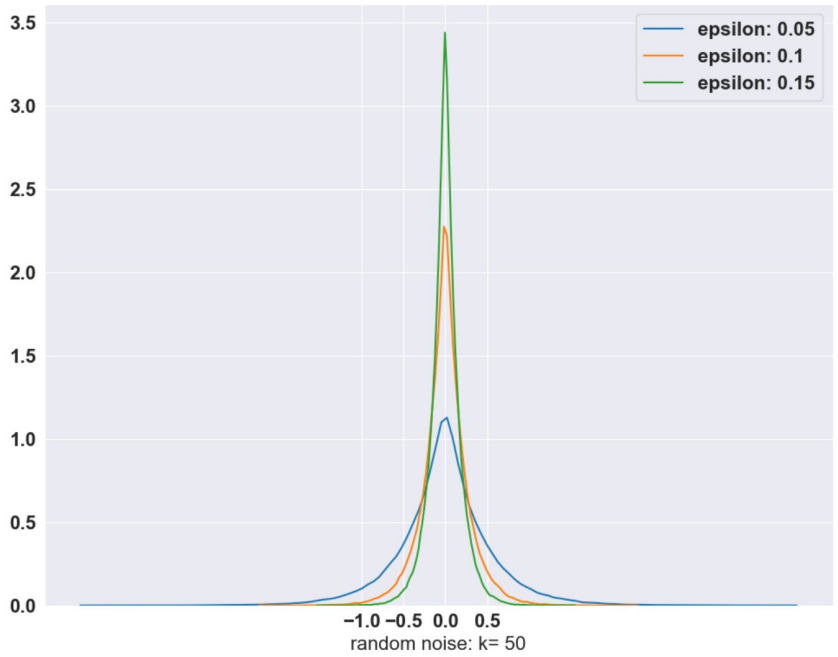
The sensitivity for the standard LR model:

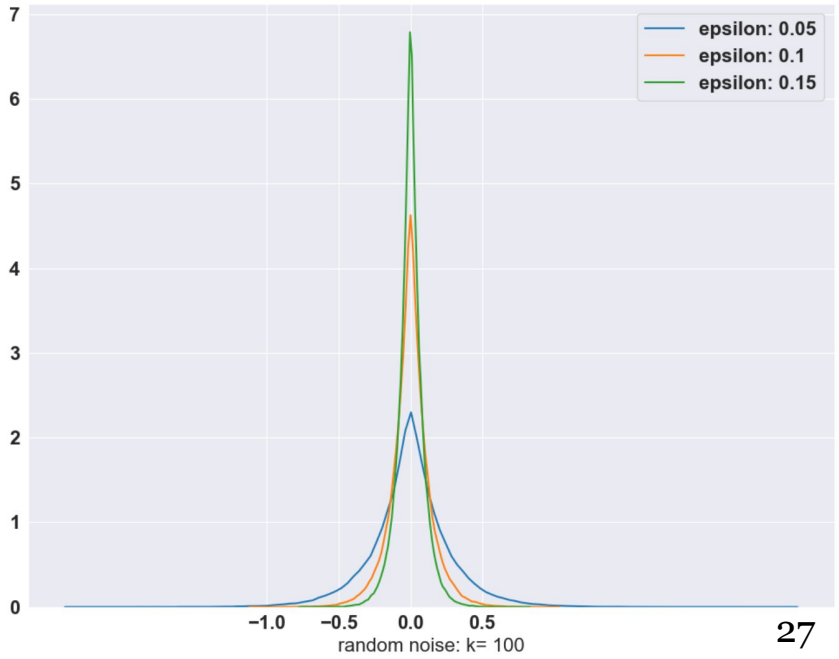
- Query the dataset for each instance's label
- The maximal difference of the query is 1 when querying the label
- Unbearable weight of noise

Differential privacy

Selection of epsilon:

- Domain of the feature matrix is [0,1]
- Domain of the label distribution is [0,1]
- Control noise domain above 90% in [-0.5, 0.5]





Prove

Laplace noise for feature matrix causes little accuracy loss.

- f_new(θ) is the loss function for distorted dataset
- Zij is random noise vector sampled from Laplace distribution
- When number of instance of each bag is big enough,
  the new loss function stays almost the same

$$f_{new}(\theta) \approx f_{new}(0) + \theta^T f'_{new}(0) = n + \theta^T \frac{1}{m} \sum_{i=1}^{n} ((P_i - \frac{1}{2}) \sum_{j=1}^{m} X_{i,j} + z_{i,j})$$

$$= n + \theta^T \frac{1}{m} \sum_{i=1}^{n} ((P_i - \frac{1}{2})(\sum_{j=1}^{m} X_{i,j} + \sum_{j=1}^{m} z_{i,j}))$$

$$As \quad m \quad \to \infty, \quad f_{new}(\theta) = n + \theta^T \frac{1}{m} \sum_{i=1}^{n} ((P_i - \frac{1}{2})(\sum_{j=1}^{m} X_{i,j} + \sum_{j=1}^{m} E(z_{i,j}))) = f(\theta)$$

Prove

Laplace noise for label distribution causes little accuracy loss.

- Odd bag from confident instances, even bag from unconfident instances i.i.d
- Mean vectors is almost the same, treat as constant
- Noise for label distribution and Noise for feature matrix is independent
- X_bar_even as the mean vector for even bag, so is the X_bar_odd

$$f_{new}(\theta) \approx f_{new}(0) + \theta^T f'_{new}(0) = n + \theta^T \sum_{i=1}^{n} ((P_i + z_i - \frac{1}{2})\bar{X}_i)$$

$$= n + \theta^T \sum_{i=1}^{n} ((P_i - \frac{1}{2}) + z_i\bar{X}_i) = f(\theta) + \theta^T \sum_{i=1}^{n} (z_i\bar{X}_i)$$

$$\sum_{i=1}^{n} (z_i\bar{X}_i) \approx \bar{X}_{even} \sum_{i=1,i\%2=0}^{n} z_i \quad + \bar{X}_{odd} \sum_{i=1,i\%2=0}^{n} z_i$$

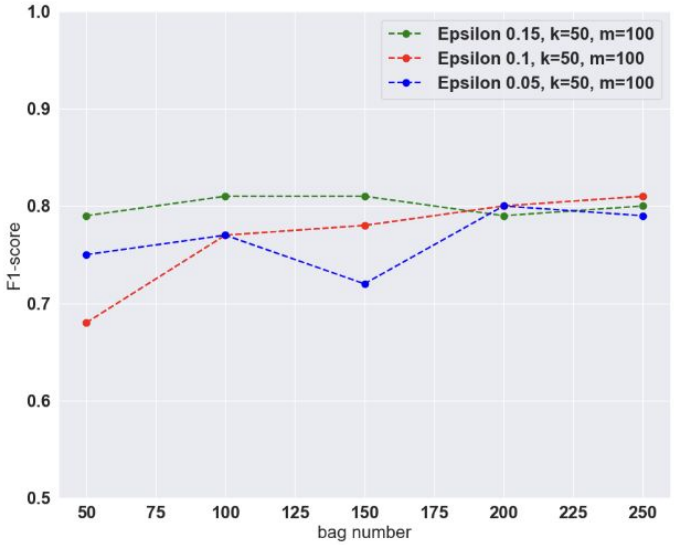
$$When \quad n \to \infty, \quad \bar{X}_{even} \sum_{i=1,i\%2=0}^{n} E(z_i) \quad + \bar{X}_{odd} \sum_{i=1,i\%2=1}^{n} E(z_i) = 0$$

$$f_{new}(\theta) = f(\theta)$$

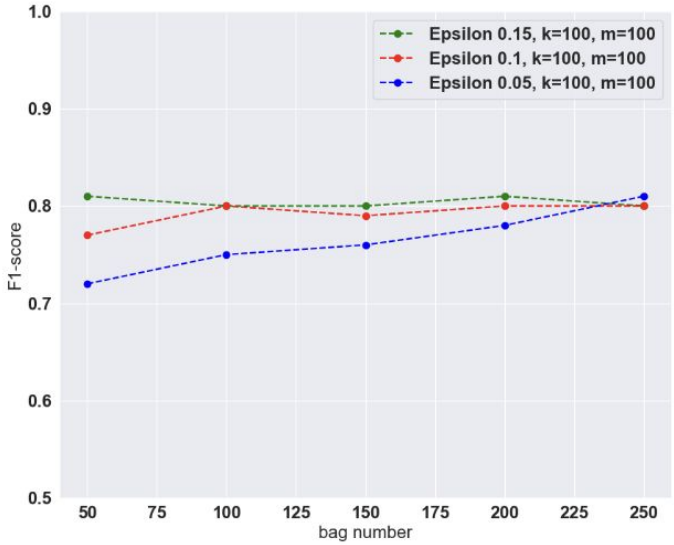
Performance

Performance increases with the number of bag increases.

- Based on Random K aggregation method
- For the dataset with less noise, the performance is more robust.
- Information loss does not change when number of bag increases
- The performance converges as the sum of the noise comes close to zero.



(a) K-value 50

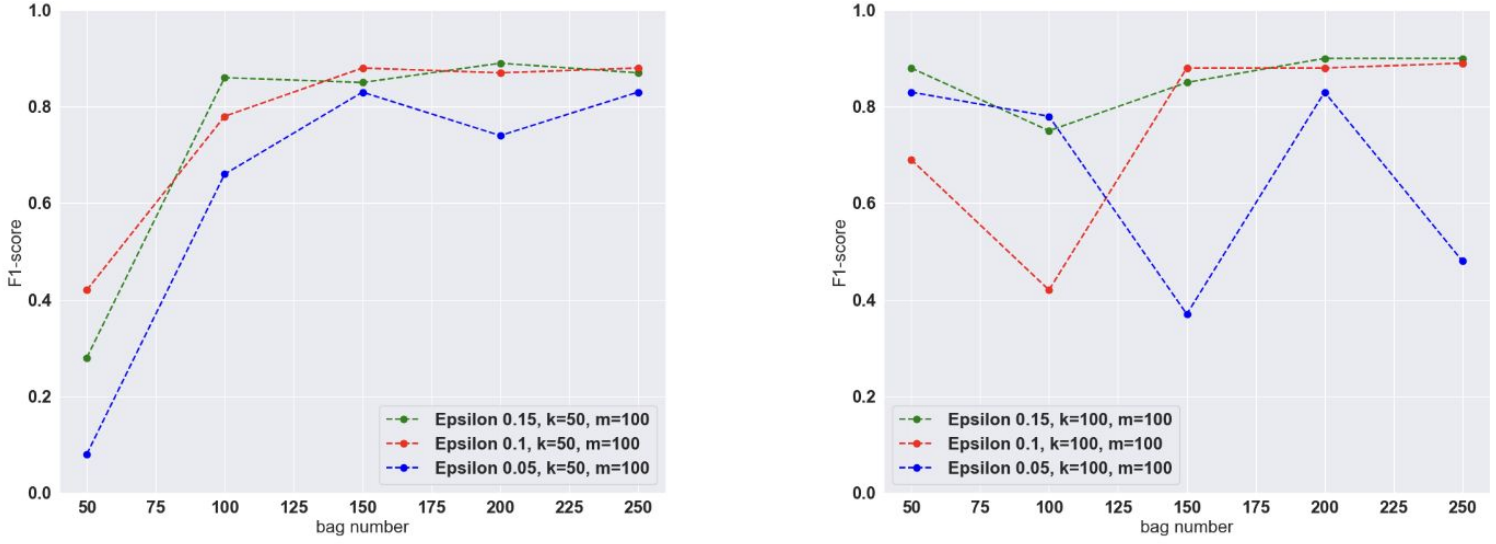


(b) K-value 100

Performance

Performance increases with the number of bag increases.

- Based on Random K aggregation method
- For the dataset with less noise, the performance is more robust.
- Information loss does not change when number of bag increases
- The performance converges as the sum of the noise comes close to zero.



(a) K-value V.S. F1-score

(b) K-value V.S. Information loss

Figure 4.6. Instagram Comment Dataset