Privacy Preserving Data Mining: Additive Data Perturbation
Outline

• Input perturbation techniques
  – Additive perturbation
  – Multiplicative perturbation
• Privacy metrics
• Summary
Benefits

- Easy to apply – applied separately to each data point (record)
- Low cost
- Can be used for both web model and corporate model
Additive noise (randomization)

Reveal entire database, but randomize entries.

User

Database

Add random noise $\varepsilon_i$ to each database entry $x_i$

For example, if distribution of noise has
Learning decision tree on randomized data

- Alice's age: 30 | 70K | ...
  - Randomizer: 65 | 20K | ...
    - Reconstruct Distribution of Age
    - Reconstruct Distribution of Salary
- Add random number to Age:
  - 30 becomes 65 (30+35)
  - Randomizer: 50 | 40K | ...
    - Reconstruct Distribution of Salary
- Classification Algorithm
- Model
Additive perturbation

• Definition
  – $Z = X + Y$
  – X is the original value, Y is random noise and Z is the perturbed value
  – Data Z and the parameters of Y are published
    • e.g., Y is Gaussian $N(0,1)$

• History
  – Used in statistical databases to protect sensitive attributes (late 80s to 90s)

• Benefit
  – Allow distribution reconstruction
  – Allow individual user to do perturbation
    • Publish the noise distribution
Applications in data mining

• Distribution reconstruction algorithms
  – Rakesh’s algorithm
  – Expectation-Maximization (EM) algorithm

• Column-distribution based algorithms
  – Decision tree
  – Naïve Bayes classifier
Major issues

• Privacy metrics
• Distribution reconstruction algorithms
• Metrics for loss of information
  – A tradeoff between loss of information and privacy
Privacy metrics for additive perturbation

- Variance/confidence based definition
- Mutual information based definition
Variance/confidence based definition

• Method
  – Based on attacker’s view: value estimation
    • Knowing perturbed data, and noise distribution
    • No other prior knowledge
  – Estimation method

  Confidence interval: the range having c% prob
  that the real value is in

  Perturbed value

Y: zero mean, std σ

Given Z, X is distant from Z in a range with c% conf
Problem with Var/conf metric

• No knowledge about the original data is incorporated
  – Knowledge about the original data distribution
    • which will be discovered with distribution reconstruction, in additive perturbation
    • can be known in prior in some applications

  – Other prior knowledge may introduce more types of attacks
    • Privacy evaluation need to incorporate these attacks
• Mutual information based method
  – incorporating the original data distribution
  – Concept: Uncertainty $\rightarrow$ entropy
    • Difficulty of estimation... the amount of privacy...

  – Intuition: knowing the perturbed data $Z$ and the noise $Y$ distribution, how much uncertainty of $X$ is reduced.
    • $Z,Y$ do not help in estimate $X$ $\rightarrow$ all uncertainty of $X$ is preserved: privacy = 1
    • Otherwise: $0 \leq$ privacy $< 1$
• Definition of mutual information
  – Entropy: $h(A) \rightarrow$ evaluate uncertainty of $A$
    • Uniform distributions $\rightarrow$ highest entropy
  – Conditional entropy: $h(A|B)$
    • If we know the random variable $B$, how much is the uncertainty of $A$
    • If $B$ is not independent of $A$, the uncertainty of $A$ can be reduced, ($B$ helps explain $A$) i.e., $h(A|B) < h(A)$
  – Mutual information $I(A;B) = h(A)-h(A|B)$
    • Evaluate the information brought by $B$ in estimating $A$
    • Note: $I(A;B) = I(B;A)$
Distribution reconstruction

• Problem: $Z = X + Y$
  – Know noise $Y$’s distribution $F_Y$
  – Know the perturbed values $z_1, z_2, \ldots, z_n$
  – Estimate the distribution $F_X$

• Basic methods
  – Rakesh’s method
  – EM estimation
Rakesh’s algorithm

• Find distribution $P(X | X+Y)$
• three key points to understand it
  – Bayes rule:
    • $P(X | X+Y) = P(X+Y | X) \frac{P(X)}{P(X+Y)}$
  – Conditional prob
    • $f_{X+Y}(X+Y=w | X=x) = f_Y(w-x)$
  – Prob at the point $a$ uses the average of all sample estimates

\[
f_X'(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X(z) \, dz}
\]
• The iterative algorithm

(1) \( f_X^0 := \text{Uniform distribution} \)

(2) \( j := 0 // \text{Iteration number} \)

repeat

(3) \( f_X^{j+1}(a) := \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X^j(a)}{\int_{-\infty}^{w_i} f_Y(w_i - z) f_X^j(z) \, dz} \)

(4) \( j := j + 1 \)

until (stopping criterion met)

Stop criterion: the difference between two consecutive \( f_X \) estimates is small
Make it more efficient...

- Bintize the range of $x$

- Discretize the previous formula

$$f'_X(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(m(w_i) - m(a)) f_X(I(a))}{\sum_{t=1}^{k} f_Y(m(w_i) - m(I_t)) f_X(I_t) L_t}$$

$m(x)$ mid-point of the bin that $x$ is in

$L_t =$ length of interval $t$
Evaluating loss of information

• The information that additive perturbation wants to preserve
  – Column distribution

• First metric
  – Difference between the estimate and the original distribution
Evaluating loss of information

• Indirect metric
  – Modeling quality
    • The accuracy of classifier, if used for classification modeling
  – Evaluation method
    • Accuracy of the classifier trained on the original data
    • Accuracy of the classifier trained on the reconstructed distribution
DM with Additive Perturbation

- Example: decision tree
- A brief introduction to decision tree algorithm
  - There are many versions...
  - One version working on continuous attributes

<table>
<thead>
<tr>
<th>rid</th>
<th>Age</th>
<th>Salary</th>
<th>Credit Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>23</td>
<td>50K</td>
<td>High</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>30K</td>
<td>High</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
<td>40K</td>
<td>High</td>
</tr>
<tr>
<td>3</td>
<td>68</td>
<td>50K</td>
<td>Low</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>70K</td>
<td>Low</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>20K</td>
<td>High</td>
</tr>
</tbody>
</table>

(a) Training Set

Age < 25

Partition(Data S)

begin
(1) if (most points in S are of the same class) then
(2) return;
(3) for each attribute A do
(4) evaluate splits on attribute A;
(5) Use best split to partition S into S1 and S2;
(6) Partition(S1);
(7) Partition(S2);
end

Initial call: Partition(Training Data)
When to reconstruct distribution

- Global – calculate once
- By class – calculate once per class
- Local – by class at each node

- Empirical study shows
  - By class and Local are more effective
Summary

• We discussed the basic methods with additive perturbation
  – Definition
  – Privacy metrics
  – Distribution reconstruction

• The problem with privacy evaluation is not complete
  – Attacks