COMP 551: Applied Machine Learning

Project 4 Presentation :

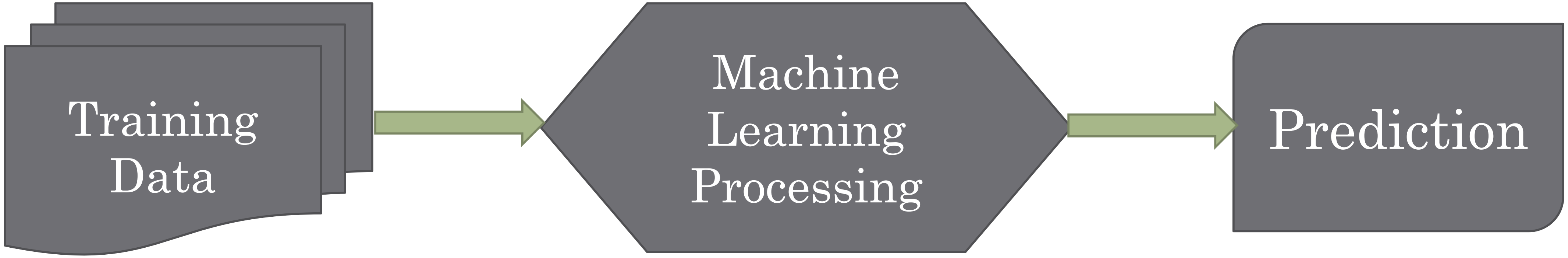
Methods for Responsible Machine Learning

Topic: Privacy Preserving for Machine Learning in Banking Applications

Hui Lee Ooi
Raylene MacDonald
Amèn Memmi

Introduction: Motivation

- Banking industry, process huge amount of data, tackle problems at large scale.
- Automated processing and decision making.
- Security



Introduction: Motivation

- Too little information: insufficient knowledge for processing the data.
- Too much information : privacy infringement
- Tradeoff?
- Maximize the amount of useful information while preventing sensitive information from being inferred.

	Gender	Age	Profession	Annual Earning
Ali	Male	34	Doctor	…
Bob	Male	55	Engineer	…
Cecilia	Female	42	Teacher	…
Dan	Male	23	Programmer	…

Introduction: Dataset

- Dataset on default of credit card in Taiwan.
- Dataset containing 30000 clients with 23 features from UCI Machine Learning Repository.
- Attribute information:
    - X1: amount of given credit
    - X2: gender
    - X3: education
    - X4: marital status
    - X5: age
    - X6-X11: history
    - X12-X17: amount of bill statement
    - X18-X23: amount of previous payment

Binary output variable: default on next bill payment (1 or 0)

SVD (Singular Value Decomposition)

- SVD is a matrix factorization method generally used for dimensionality reduction; however, it can also be used for data distortion.

The singular value decomposition of the matrix A is [14]

$$A = U\Sigma V^T,$$

where U is an $n \times n$ orthonormal matrix, $\Sigma = \mathrm{diag}[\sigma_1, \sigma_2, \ldots, \sigma_s]$ $(s = \min\{m, n\})$ is an $n \times m$ diagonal matrix whose nonnegative diagonal entries are in a descending order, and V^T is an $m \times m$ orthonormal matrix. The number of nonzero diagonals of Σ is equal to the rank of the matrix A.

[1]

- We choose the k largest diagonals of Σ, representing the highest variance among the attributes, to obtain $A_k = U_k\Sigma_k V_k$, an n x k matrix approximating the original n x m data matrix A.

Impact of SVD on Classification Accuracy

- Using scikit-learn's Random Forest classifier on the original data, we obtained, through 5-fold cross-validation, an accuracy of 0.81. We then experimented with different values of k for SVD.

Data distortion	Accuracy
Untouched data	0.81
SVD with k = 15	0.79
SVD with k = 10	0.76
SVD with k = 5	0.73

- Even for small values of k, the accuracy remains reasonable. This indicates the output variable is closely correlated to a small subset of the original 23 features.

Data Perturbation with Random Noise

- We normalize the values of our data matrix A to the [0, 1] range
- We compute $A_u = N_u + A$, where the values of N_u are uniformly distributed in the [0, 1] range
- We compute $A_n = N_n + A$, where the values of N_n are normally distributed with parameters μ and σ

Data distortion	Accuracy of Random Forest Classifier
Original, normalized data	0.81
With uniform noise from [0, 1]	0.78
With Gaussian noise, $\mu = 0$ and $\sigma = 0.1$	0.76
With Gaussian noise, $\mu = 0$ and $\sigma = 0.5$	0.75
With Gaussian noise, $\mu = 0.5$ and $\sigma = 0.5$	0.74

How secure are these methods?

We use the Value Difference metric to measure how much the original data, A, has been altered to yield the distorted data matrix A' :

$$VD = \| A - A' \| / \| A \|$$

Data Distortion	Value Difference
SVD with k = 15	0.001
SVD with k = 10	0.049
SVD with k = 5	0.170
With uniform noise from [0, 1]	2.203
With Gaussian noise, $\mu = 0$ and $\sigma = 0.1$	0.393
With Gaussian noise, $\mu = 0.5$ and $\sigma = 0.5$	2.625

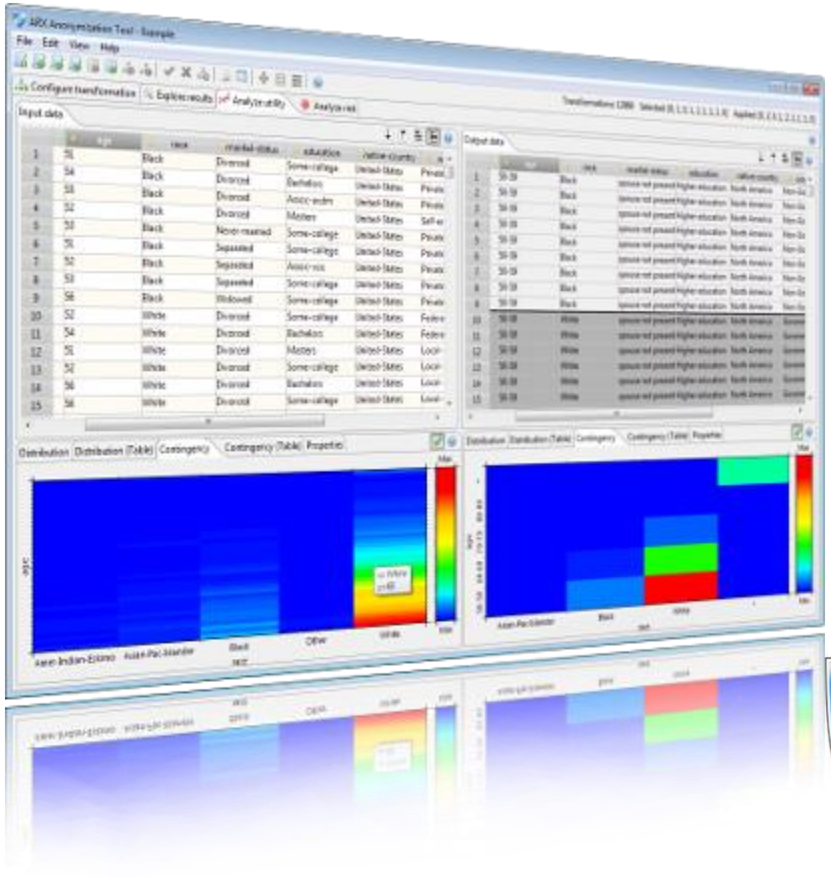
A suggested [1] Value Difference of at least 0.15 is deemed sufficient for privacy-preserving methods; however, this depends on the domain.

K-Anonymity

- Attributes are altered until each row is identical with at least k-1 other rows:
  - *Suppression* – can replace individual attributes with a *.
  - *Generalization* – replace individual attributes with a broader category according to a certain hierarchy. Example: (Age: 26 => Age: [20-30]; Occupation: Police officer => Governmental)
- K-Anonymity thus prevents definite database linkages. At worst, the data released narrows down an individual entry to a group of k individuals
- Unlike Output Perturbation models, K-Anonymity guarantees that the data released is accurate

Implementation

- ARX anonymization tool: open source software used to anonymize data
- We defined hierarchy for each attribute
- ARX scans solution space and chooses optimal Suppression – Generalization combo to minimize risk.



Input vs Output Data

Input data | Classification accuracy

	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2
1	92082	120077	92593	168019	5000
2	73310	240738	135722	304815	8000
3	61351	126198	124746	168096	6409
4		0	0	0	0
5	76	3551	0	2731	757
6	5	326	-87	0	0
7	211	9422	8853	0	0
8	7585	18301	18663	390	780
9		9263	3693	7244	4000
10	938	0	0	1939	2040
11		0	0	0	0
12		0	0	0	0
13	1000	-1000	-1000	0	0
14	242	8123	8993	0	0
15		0	0	0	0
16		0	0	0	0
17	400	0	0	0	0
18		0	0	0	0
19		0	0	0	0
20		0	0	0	0
21		0	0	0	0
22		0	0	0	0
23		1864	1720	0	0
24		400	0	5275	0

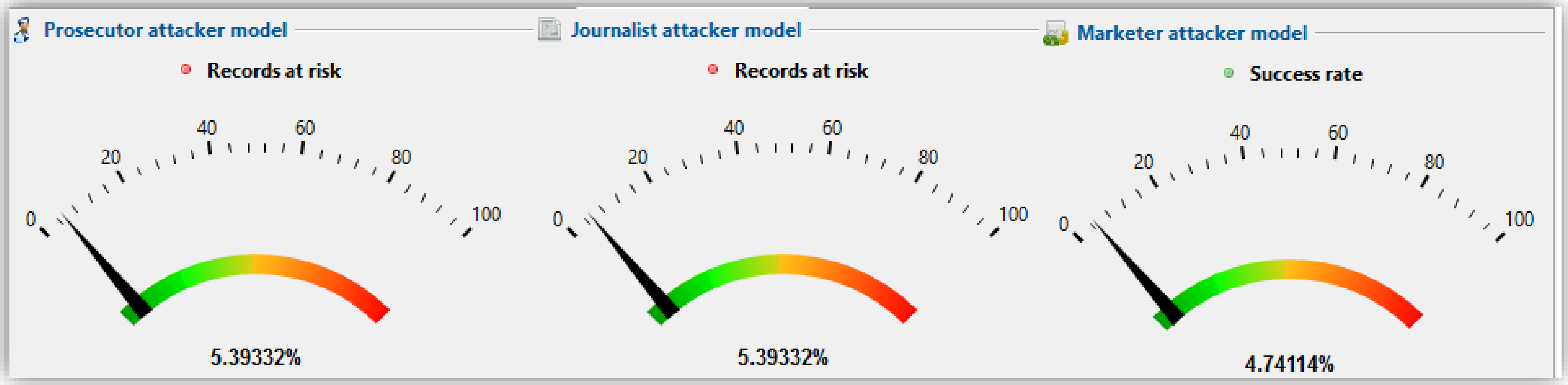
Output data | Classification accuracy

	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2
1	50000, 270000[	120077	92593	>=160000	*
2	50000, 270000[	240738	135722	>=160000	*
3	50000, 270000[	126198	124746	>=160000	*
4	50000, 270000[	0	0	[0, 160000[	*
5	50000, 270000[	3551	0	[0, 160000[	*
6	50000, 270000[	326	-87	[0, 160000[	*
7	50000, 270000[	9422	8853	[0, 160000[	*
8	50000, 270000[	18301	18663	[0, 160000[	*
9	50000, 270000[	9263	3693	[0, 160000[	*
10	50000, 270000[	0	0	[0, 160000[	*
11	50000, 270000[	0	0	[0, 160000[	*
12	50000, 270000[	0	0	[0, 160000[	*
13	50000, 270000[	-1000	-1000	[0, 160000[	*
14	50000, 270000[	8123	8993	[0, 160000[	*
15	50000, 270000[	0	0	[0, 160000[	*
16	50000, 270000[	0	0	[0, 160000[	*
17	50000, 270000[	0	0	[0, 160000[	*
18	50000, 270000[	0	0	[0, 160000[	*
19	50000, 270000[	0	0	[0, 160000[	*
20	50000, 270000[	0	0	[0, 160000[	*
21	50000, 270000[	0	0	[0, 160000[	*
22	50000, 270000[	0	0	[0, 160000[	*
23	50000, 270000[	1864	1720	[0, 160000[	*
24	50000, 270000[	400	0	[0, 160000[	*

Anonymity gain – Identification risk reduction

<u>Before 3- anonymity:</u>

<u>After 3- anonymity:</u>



Effect on ACC and AUC

K-anonymity order	Original data	K=2	K=5	K=10	k=20	K=50	K=100	K=500	K=1000
Logistic Regression	81.06	80.16	80.77	79.74	78.12	78.44	79.12	78.45	78.44
Naïve Bayes	80.93	80.43	81.2	79.56	79.56	79.76	79.5	79.26	78.81
Random Forest	82.16	81.13	80.63	80.63	80.53	79.46	80.2	78.26	78.33

Accuracy(ACC) vs K-anonymity order K for 3 classifiers

K-anonymity order	Original data	K=2	K=5	K=10	k=20	K=50	K=100	K=500	K=1000
Logistic Regression	61.14	60.55	60.75	60.12	59.88	59.74	59.48	60.15	60.01
Naïve Bayes	72.90	71.92	73.75	70.05	70.56	70.76	70.65	70.94	70.66
Random Forest	77.01	75.01	72.70	72.35	72.12	70.40	72.92	66.79	66.5

Area Under Curve (AUC) vs K-anonymity order K for 3 classifiers

References

1. S. Xu, J. Zhang, D.Han, J.Wang, Data Distortion for Privacy Protection in a Terrorist Analysis System. University of Kentucky, 2005.
2. S. Han, W. Keong Ng, P. S. Yu, Privacy-Preserving Singular Value Decomposition. IEEE International Conference on Data Engineering, 2009
3. M. Naga Lakshmi1 & K Sandhya Rani, SVD based Data Transformation Methods for Privacy Preserving Clustering. International Journal of Computer Applications Volume 78 – No.3, September 2013.