IBM Research: Mobile, Solutions, and Mathematical Sciences

Health Insurance Market Risk Assessment:
Covariate Shift and k-Anonymity

DENNIS WEI, KARTHIKEYAN NATESAN RAMAMURTHY, AND KUSH R. VARSHNEY

A Tale of Two Laws

PATIENT PROTECTION AND AFFORDABLE CARE ACT

- Changed the landscape of the health insurance market in the United States
- Health insurance companies had to decide which new markets to enter
  - New markets defined by geography, age group, and other prospect base criteria

HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT

- Required all releases of health-related information about individuals to protect their privacy
  - Even for health insurance companies' internal planning uses
- k-Anonymity is a common mathematical interpretation of the privacy condition

Outline

- **The Market Risk Assessment Regression Problem**
- Three Population Shift
- k-Anonymity for Workloads Requiring Distribution Preservation
- Empirical Results
- Summary

Desiderata of Health Insurance Companies

- Desire low-cost (healthy) people enroll in their plans

- Not allowed to accept or deny enrollment on an individual basis
- Allowed to offer or not offer plans in well-defined markets
- (Allowed to have marketing strategies)

- Use data-driven decision making for determining whether or not to offer plans in new markets

Desiderata of Health Insurance Companies

- Desire cost data on people who will enroll in new markets

- Only have cost data on people who have enrolled in existing markets
- Have demographic data on existing market
- Have demographic data on new market

- Regression problem with covariate shift
  - Also need to consider enrollment: three-population shift

Demographic and Cost Data Availability

	Existing Market	New Market
Enrolled	insurance company has demographic data and cost data	insurance company has **no** demographic data or **cost data**
Everyone (enrolled and not enrolled)	insurance company can get demographic data from public sources	insurance company can get demographic data from public sources

The Market Risk Assessment Regression Problem

- Use cost and demographic data for enrolled members in the existing market, demographic data for the existing market, and demographic data for the new market to estimate cost for enrolled members in the new market
- Can use regression technique of choice
  - Ordinary least-squares with and without log-transformed data
  - Two-part models
  - Generalized linear models
  - Multiplicative regression
- Need a type of covariate shift to account for the differences between the features of the existing and new market

Outline

- The Market Risk Assessment Regression Problem
- **Three Population Shift**
- k-Anonymity for Workloads Requiring Distribution Preservation
- Empirical Results
- Summary

Regression Problem with Covariate Shift

- Predict cost of individual from his or her demographic features
- Let Y be the cost and X be the demographic features
- Training samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ drawn from $p_{X,Y} = p_X p_{Y|X}$
- Test samples drawn from $q_X p_{Y|X}$ where q_X is a different feature distribution
  - Covariate shift
- Minimize risk weighted by importance of each sample under q_x
- $\widehat{Y}(\cdot) = \arg\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} \frac{q_X(x_i)}{p_X(x_i)} L\left(f(x_i), y_i\right)$

Taking Enrollment Into Account

- Let $M = 0$ indicate the existing market and $M = 1$ indicate the new market
- Let $E = 0$ indicate a non-enrolled person and $E = 1$ indicate an enrolled person

- The training feature distribution is $p_{X|E,M}(x \mid e = 1, m = 0)$
- The testing feature distribution is $p_{X|E,M}(x \mid e = 1, m = 1)$
  - This cannot be estimated directly
  - Different than the typical covariate shift problem
  - Cannot simply use an importance weight of $p_{X|E,M}(x \mid e = 1, m = 1) \,/\, p_{X|E,M}(x \mid e = 1, m = 0)$

Three Population Shift

- Demographic distributions for the existing and new markets are available: $p_{X|M}(x \mid m = 0)$ and $p_{X|M}(x \mid m = 1)$

- Use Bayes theorem to obtain:

$$\frac{p_{X|E,M}(x|1,1)}{p_{X|E,M}(x|1,0)} \propto \frac{p_{E|X,M}(1|x,1)}{p_{E|X,M}(1|x,0)} \frac{p_{X|M}(x|1)}{p_{X|M}(x|0)}$$

- Assume that the probability of enrollment conditioned on features and market is independent of the market once the features are fixed: $p_{E|X,M}(1 \mid x, m) = p_{E|X}(1 \mid x)$

Three Population Shift

- Then:

$$p_{X|E,M}(x|1,1) \propto p_{X|E,M}(x|1,0)\frac{p_{X|M}(x|1)}{p_{X|M}(x|0)}$$

- and:

$$\hat{Y}(\cdot) = \arg\min_{f \in F} \frac{1}{n}\sum_{i=1}^{n} \frac{p_{X|M}(x_i|1)}{p_{X|M}(x_i|0)} L(f(x_i), y_i)$$

- Estimate importance weight non-parametrically or with logistic regression

Outline

- The Market Risk Assessment Regression Problem
- Three Population Shift
- **k-Anonymity for Workloads Requiring Distribution Preservation**
- Empirical Results
- Summary

k-Anonymity

- Mathematical standard for adhering to HIPAA when doing microdata release

- 3 types of variables
  1. Identifiers, e.g. name
     - Drop immediately for anonymization
  2. Quasi-identifiers, e.g. gender, birthdate, postal code
     - Can be used to match with other data that has both identifiers and quasi-identifiers, e.g. voting roll
  3. Sensitive data, e.g. health data
- Under k-Anonymity, change the data so that the quasi-identifiers for an individual cannot be distinguished from at least $(k-1)$ other individuals

Workload Determines Quality of Anonymization

- To achieve k-anonymity, we can group the records so that the smallest group has at least k elements
- Any such grouping is equally good from the privacy perspective
- Quality of the data transformation depends on the workload for which the data will be used afterwards
- For the three population shift and regression workload, we would like transformation to preserve the probability distribution of the data

k-Member Clustering

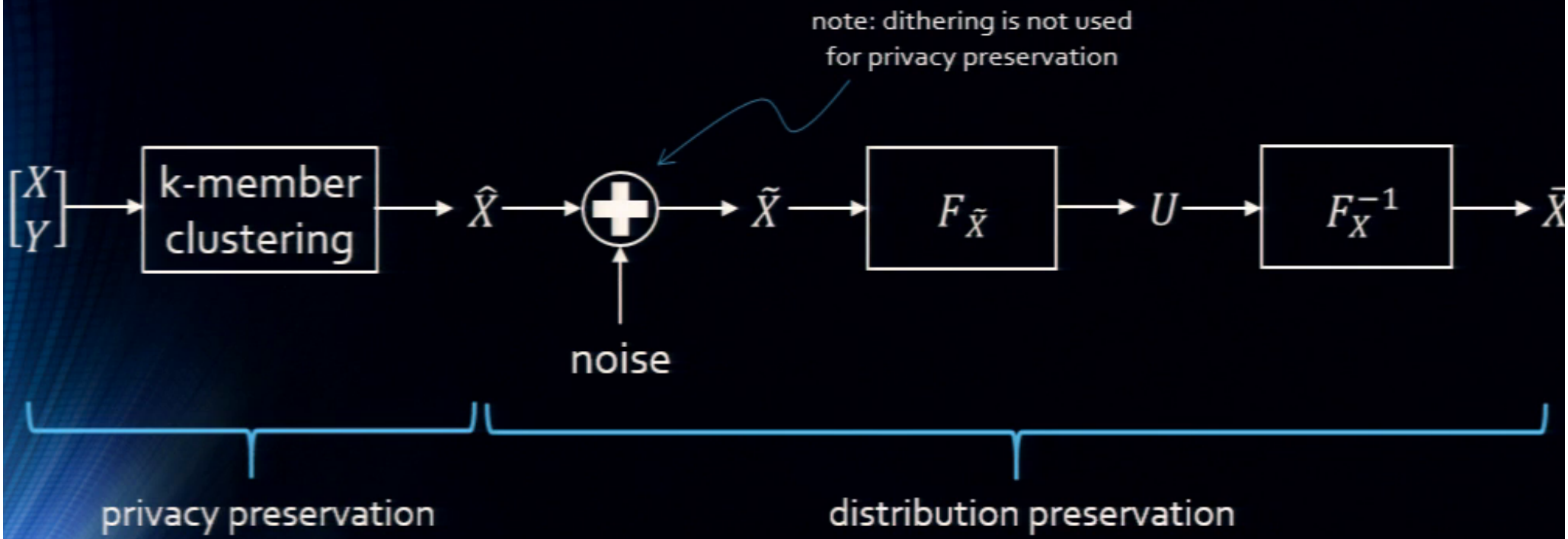
- Clustering is a good way of grouping records to achieve k-anonymity
- Most clustering algorithms, including k-means, take the number of clusters as a parameter, not the minimum cardinality of the clusters
- k-Member clustering is a method that does allow the specification of that parameter
- The centroids or representation points obtained by k-member clustering do not follow the probability distribution of the original data

Distribution-Preserving Quantization

- Distribution-preserving quantization is an alternative to k-means to allow the resulting data to follow the distribution of the original data
  - Based on subtractive dithered quantization + Rosenblatt's transformation
  - Developed for audio signals
- Specification is still on the number of clusters, not on minimum cluster size

- We propose a distribution-preserving k-member clustering
  - k-Anonymization matched to the covariate shift workload

Block Diagram



Outline

- The Market Risk Assessment Regression Problem
- Three Population Shift
- k-Anonymity for Workloads Requiring Distribution Preservation
- **Empirical Results**
- Summary

Health Care Data

- We have worked with health insurance companies in the recent past, but their data is confidential
- Empirical results on data from the Medical Expenditure Panel Survey
- MEPS is not from an insurance company so data does not have notion of enrollment and markets
  - Treat overall United States as existing market and 'rating areas' in California as new markets
  - Model plan enrollment using true Obamacare enrollment distribution
- Demographic data from American Community Survey
- Gender, age (binned), education level, income level

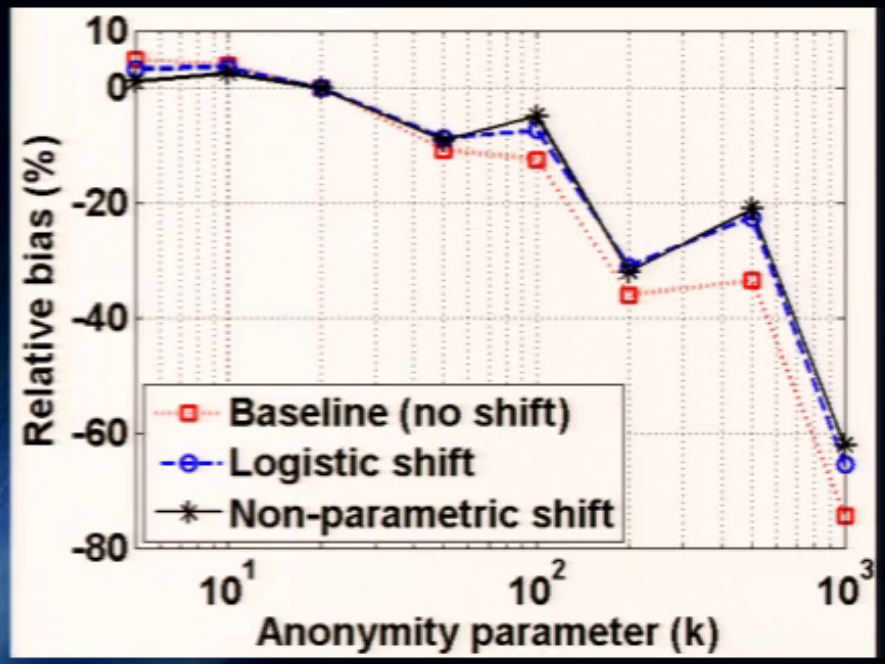
Regression and Performance Evaluation

- All demographic features in our data set are discrete
- Use linear combination of nonlinear univariate functions of each demographic feature as the function class for the regression
  - (Advanced regression is not the focus of this paper)

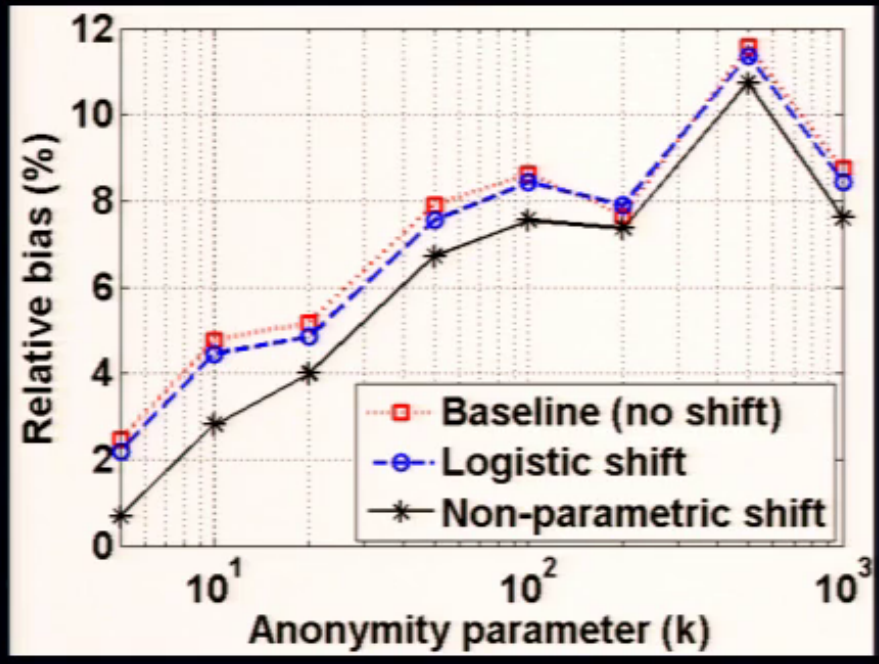
- Most important performance metric is the bias of the prediction
  - Paper also has R^2 results

Comparison of Anonymization Techniques: Relative Bias

Conventional k-Member Clustering



Distribution-Preserving k-Member Clustering



Outline

- The Market Risk Assessment Regression Problem
- Three Population Shift
- k-Anonymity for Workloads Requiring Distribution Preservation
- Empirical Results
- **Summary**

Summary

- Addressed the health insurance market risk assessment problem
  - Especially pertinent after passage of Affordable Care Act
- Insurance companies haven't used sophisticated data mining and machine learning approaches for estimating costs in new markets
- Only enrolled member cost data in existing market available
- Cost prediction in new markets requires novel three population shift
- HIPAA requires privacy preservation
- Distribution-preserving k-member clustering
- Excellent empirical performance