January 8 - Progress

**Literature Review**

***The Limits of Current Information Loss Metrics as a Measure for Data Utility in k-Anon***

Imperial Project last year – Thomas Marchand.

Useful information contained:

* Formal definitions of Generalizations
* Hierarchy based metrics, and others
* Introduces Entropy
* Introduction to Datafly, MinGen (pseudocode)
* Introduction to Incognito
* Introduction to Mondrian (pseudocode)
* Methodology
  + Auto-sklearn to hyper parameter tune automatically
* Result: mixed ; question on datasets

***k-ANONYMITY: A model for Protecting Privacy***

First paper on k-anon – Latanya Sweeney.

Useful information contained:

* Justification for k-anon
  + Risks on un-anonymized datasets
  + Formal basic definitions
* Weaknesses of k-anon

***Achieving k-anonymity Privacy Protection Using Generalization and Suppression***

Follow up on first paper; how to actually k-anon – Latanya Sweeney.

Useful information contained:

* Formal definitions of Generalization
  + Value Generalization Hierarchy
* Formal definitions of Suppression
* Metrics and “minimal distortion” when anonymizing
* MinGen (pseudocode)
  + Distorts minimally but inefficient
* Datafly (pseudocode)
  + Efficient but more distortion
  + Makes decisions at tuple level level (crude) whereas MinGen generalizes cells individually
* µ-Argus (pseudocode)

**Search for a Dataset**

**Contraceptive Method Classification**

Origin: 1987 National Indonesia Contraceptive Prevalence Survey

Problem: predict the current contraceptive method choice

Attributes: 9 categorical attributes

Size: 1473

**Post-Operative Life Expectancy Classification**

Origin: Wroclaw Thoracic Surgery Center on patients who underwent major lung resections for lung cancer

Problem: predict patient’s survival a year after the operation

Attributes: 17 categorical or numerical attributes

Size: 470

**Adult Census Salary Classification**

Origin: 1994 US census

Problem: predict if income is superior to 50K

Attributes: 14 attributes

Size: 48842

**Adult – Hyperparameter tuning**

Chose Adult because it is a large set with a good range of attributes. The post-operative life expectancy classification is too small. I’m not entirely discounting the contraceptive set entirely and will probably use it later too.

Before automating all the hyperparameter tuning, I manually tuned a classifier



Best accuracy on a training set (CV=5): ~ 84%

**Questions**

* Should I look into alternative Data Science Methods (Random Forests, SVMs…)
  + Random forests split on attributes so would be interesting to find out how generalizing these attributes affects the accuracies.
* Removing columns and testing subsets of attributes