Data Mining Process

SYS6018, Matthew S. Gerber

# Ground Rules

This document outlines a process that can be applied to data mining problems. Please adhere to the following rules when working through the steps in the process.

* **No jargon:** An example of jargon is “cloud computing”. This concept has many meanings including distributed storage and/or computing, virtualized architectures, service-oriented architectures, and many others. Replace jargon with the technical concepts that you actually have in mind.
* **Use clear, concise, and complete sentences:** Enough said.
* **Use mathematical notation when doing so is helpful:** Mathematical notation is often used to great advantage when expressing complex ideas concisely. It is also often used to impress, confuse, and abuse the reader. Be judicious.
* **Write for a technical audience:** You should write for a technical audience that already understands the concepts you are writing about.
* **Smooth it over:** The entire team should contribute to this document, but one individual should be responsible for smoothing the document into a coherent whole. The document should speak with one voice.
* **Cite references appropriately:** You may use any citation format you wish, but all concepts and claims should be appropriately referenced.

The actual process follows.

# 1. Identify the Problem

Before applying data mining we must first clearly identify the problem that we aim to address and the envisioned impact of addressing this problem.

#### Descriptive Analysis

* What is the current situation with respect to the problem? Provide examples of the issues that need to be addressed.
* What are the surrounding events and other factors that may be contributing to or correlated with the current situation?
* What data exist that describe the current situation?

#### Normative Analysis

* What is the desired alternative situation? Provide examples that demonstrate the desired improvement.

#### Stakeholders

* Which individuals, businesses, or other agencies care about this problem and why?

#### Impact

* What is the envisioned impact of transitioning from the current to the desired alternative situation?

# 2. Define Objectives and Metrics

#### Objectives

* What are you trying optimize, increase, decrease, or otherwise change in order to transition to the desired alternative situation?

#### Metrics

* How will you measure these changes?

# 3. Understand the State-of-the-Art

* What data sources have been investigated?
* What data mining methods have been applied to the data?
* How were the methods evaluated and what were the evaluation outcomes?
* Given what others have already done, why does the problem remain unsolved? What makes it difficult?

# 4. Define Hypotheses and Approach

#### Hypotheses

* What are your hypotheses?

Good hypotheses have the following characteristics.

* Predictive: When applied, a hypotheses makes a prediction about what will happen.
* Testable in a finite time: It is possible, given a finite amount of time, to execute a test that will definitively support or oppose the hypothesis.
* Falsifiable: It is possible for the hypothesis to be opposed by the test.
* Novel: The hypothesis has not been investigated in the past.
* Lead directly to data, methods, and experiments: The hypothesis leads directly to an experimental design.

Example hypotheses:

* Drug X will produce better recovery rates than Drug Y when administered to patients.
* Data mining method X will produce a better ROC curve than method Y when applied to data Z.
* Adding predictor A to model M will improve ROC performance.

#### Data

* What process led to the finalized data that you have obtained for this problem?
* What biases might exist in the data?

#### Methods

* What data mining methods are you applying to the data?
* Are the assumptions of your methods met?
* Are your methods novel compared to prior related work? If so, how?

#### Evaluation Setup

* How will you support or oppose your hypotheses using measurements obtained from your data and methods?
* Does your evaluation setup faithfully measure what you intend?

# 5. Execute Approach and Report Results

* What did you learn about the problem?
* What did you learn about the data, methods, and evaluation?
* Do your results support or oppose your hypotheses?
* How does your work help to move us closer to the desired alternative situation (normative scenario)?