Final Project Schema

Data Mining (4740-6740) – Fall 2017
Georgia State University
Instructor: Azim Ahmadzadeh
cs.gsu.edu/~aahmadzadeh1
Announced on: [22 Oct. 2017]
Due Date: [19 Nov. 2017] by [11:30 pm]

* PRIZE *

In addition to the grade you will get for your final project, you can get a prize:

- Top 3 works (+10/100),
- 4^{th} and 5^{th} works (+5/100),

that will be added to your final exam grade.

1. CHOOSE YOUR DATA.

Choose a dataset that interests you. You can surf the Internet and find something that tickles your curiosity. As long as it is complex enough to serve a data mining knowledge discovery, I am fine with it.
You want some good sources of data? See the list below:

- UMI Repository [http://archive.ics.uci.edu/ml/index.php]
- GOV [https://www.data.gov/]
- SNAP [http://memetracker.org/]
- HIFLD [https://hifld-dhs-gii.opendata.arcgis.com/]
- A list of data repositories: [kdnuggets/datasets/...]
- Another list of data repositories:
- [/github/...]

2. STUDY YOUR DATA.

If you carefully study all the aspects of the dataset, you will save a lot of time later. In fact, **the secret of being able to plan a reliable and realistic project** relies in how well you studied the data. So,

- try to understand each attribute.
- try to find out what knowledge you can extract from the data. Set your goal.
- try to see which parameters seem more important and which ones are not.
- try to detect the correlated parameters and handle them.
- try to find out the missing, inaccurate, erroneous values.

3. PROBLEM FORMULATION

It is very important to have a clear statement of the problem you want to answer. **The formulation of your problem MUST be clearly visible in your final poster.**
It is a great idea to formulate the "given" and "goal" parts of your problem in a mathematical form.

4. CHOOSE A MODEL

At this stage, you know your data and your problem. Based on all you have learned in our class, decide which data mining model can answer your problem better.
Ask yourself:

- Is this a rule mining task?
- Do I want to cluster data?
- Is this a classification task?
- Is there any prediction involved?
- Am I interested in ourliers?

5. PREPARE YOUR DATA

Now that you know what your data mining model is, it's time for some data preprocessing. Do you remember that the preprocessing task usually eats up about 80% of your time? Be prepared for that!
This is where you want to:

- take care of the dirtiness of your data. (missing values, inaccurate data, ⋯)
- change data types such that the data serve your task better.
- drop the features you are not interested.
- create new features based on the existing ones.
- transform your data to a new space if it could make your task easier.
- normalize your data. (Almost always you need this step.)

6. CHOOSE YOUR VALIDATION METHOD

Before you start building your model, you need to decide about your validation method.

- If your data is large enough, go with **Holdout** or **Random Subsampling**.
- If you don't have access to a large dataset, a variation of **Cross-Validation** may help you.

After you figured out what your validation method would be, divide your data into training and test set and **be very careful** that it is very tempting to use your test set while you are training your model. This would destroy all your hard work, and give you unrealistic results.

7. TEST YOUR MODEL

It's time to tell us how good your model was while you were training it, and how good it performs on the test data.
Depending on your model, there might be some parameters that you would want to report the effectiveness of your model while changing them. For instance:

- the depth of the decision tree,
- the value of k in a kNN classification model.

Remind yourself our conversation about the different evaluation measures, and why they can be misleading. Which one(s) should you use?

8. KNOWLEDGE

Finally, tell us what you discovered. This is the ultimate goal of all you tried on a data set. This is how you can sell your work. So, please, be clear, loud, and reasonable, in the final conclusion. Tell us what knowledge you have discovered from the data set you worked on.

A. YOUR POSTER

Before you start your project, take a quick look at **this amazing set of slides** to know what your final output (poster) should look like.
Find a nice templates online. [1][2]

Poster size: 24X36 (in inches)
Poster format: PDF

You need some illustrations:

- To present the subject you are studying. (e.g., Use a picture related to vaccination if you are working on a dataset concerning vaccination.)

You need to draw plots for:

- The data size (both dimension & volume)
- Type and range of each attribute.
- Distribution of data and outliers (Use box plots)
- Distribution of data after cleaning the data and normalization. (box pots)
- Test/Training data separation.
- Performance on training data.
- Performance on training data after any modifications.
- Performance on test data.