

CIDM 6355 Learning Activity 5 Instruction

Purposes: This learning activity helps you find and formulate a classification task and then discuss with your teammates.

1. Explore Common DM Tasks

Some common topics include customer churn prediction, fraud detection, revenue prediction, customer classification, price prediction, risk prediction, etc. You may visit the following three sites to get a sense of what data mining projects look like:

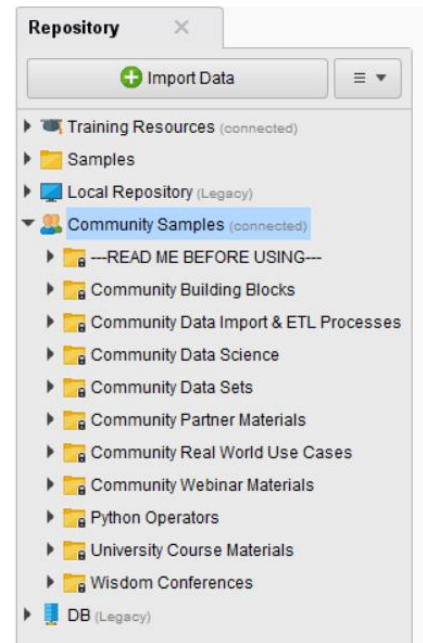
- <https://www.kaggle.com/competitions> (professional DM competitions)
- <http://archive.ics.uci.edu/ml/datasets.html> (mainly academic DM research projects)
- <http://www.galitshmueli.com/student-projects> (student DM projects)

2. Select A Dataset For Classification

Usually, we have a problem first and then find datasets to solve it. Here, we use an opposite method: having the dataset first and then formulate a problem. Please explore any of the following data sources and then find a dataset for a classification problem that is of interest to you. Select a dataset (you must include the URL) and then briefly describe it (what is it about, how and when is it obtained, how many attributes and records are included). Requirement: Your data must include at least **1,000** records and **ten** attributes.

- <https://archive.ics.uci.edu/ml/datasets.php> (Highly recommended; you can filter classification tasks).
- <https://www.kaggle.com/datasets>
- <https://wrds-www.wharton.upenn.edu/register/> (You need to register with your WTAMU email address and then your request may be approved by the WTAMU administrator).
- <http://databank.worldbank.org/data/reports.aspx?source=world-development-indicators>
- <https://www.springboard.com/blog/free-public-data-sets-data-science-project/>
- <http://www.inf.ed.ac.uk/teaching/courses/dme/2012/datasets.html>
- [Health Data Sets](#)
- <https://www.r-bloggers.com/datasets-to-practice-your-data-mining/>
- <http://data.gov>
- <http://applieddatamining.blogspot.com/>
- <https://www.basketball-reference.com/>
- <http://rotoguru1.com/cgi-bin/hyday.pl?game=fd>
- <http://www.boxofficemojo.com/>
- <http://www.rdatamining.com/resources/data>
- <https://vincentarelbundock.github.io/Rdatasets/datasets.html>
- <http://www.city-data.com/>
- <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>
- <http://www.nfl.com/stats>

- <https://www.medicare.gov/>
- <https://r-dir.com/reference/datasets.html>
- https://dreamtolearn.com/ryan/1001_datasets
- <http://appliedpredictivemodeling.com/data/>
- <https://bigml.com/user/francisco/gallery/datasets>
- Community Samples on RapidMiner community (see the screenshot in the right)
- <http://www.scan-support.com/help/sample-data-sets> (not available now)
- Self-owned or company-owned data (the data you collected or lawfully available to you)
- Other sources (must have the permission from the instructor)



3. Formulate Your Classification Problem


Then, formulate a classification problem. A sample problem could be “how to detect whether a customer is going to churn using their historical transactions?”, “how to classify documents in different categories based on their characteristics”, “how to recognize spam email by learning the characteristics of what constitutes spam vs non-spam email”, or “how customers can be classified into different categories based on their buying patterns, web store browsing patterns etc.” All these classification questions include the three elements:

- A categorical target attribute
- A classification verb (e.g., predict, classify, etc.)
- Information used for classification (e.g., on their buying patterns)

Therefore, make sure your classification problem includes the three elements above.

4. Make Your Post

Please find LA5: Group Discussion under Lessons – Group Work

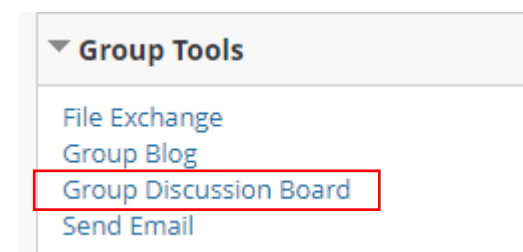

LA5: Group Discussion

Please follow LA5 Instruction to:

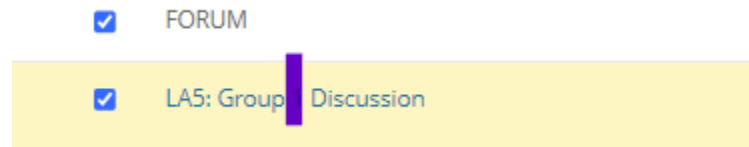
1. Make your post to share your classification problem with an appropriate dataset by creating a thread (due 9/24)
2. Make comments (e.g., asking question, providing suggestion, etc.) on the classification problem posted by your teammates (due 9/30)
3. Respond to comments made your teammates on your classification problem (e.g., answering questions asked by your teammates, providing clarification, updating your problem formulation, etc.) (due 9/30)

Each individual member should create a post to share his/her classification problem with a relevant dataset.

Once you enter into your group, please find Group Discussion Board and click it.



Then, you will find LA5: Group X Discussion. Click it and then create a thread to make your post.



When present your post, please create a thread to state your classification problem first and then describe your data.

<p>Your Problem Definition – Business Understanding</p> <p>Your Dataset – Data Understanding</p>

5. Engage Your Team

- Make comments (e.g., asking question, providing suggestion, etc.) on the classification problem posted by your teammates
- Respond to comments made your teammates on your classification problem (e.g., answering questions asked by your teammates, providing clarification, updating your problem formulation, etc.)

Grading Rubric

- Your post: Classification problem (10 points, due 9/24) must include two parts:
 - Your Problem Definition (5 points)
 - You problem must state three elements correctly
 - Your problem be clearly written
 - Data Description (5 points)
 - Your dataset should be able to answer the classification problem you formulate.
 - Your description must clearly include what is it about, how and when is it obtained, how many attributes and records are included. If any information is unavailable, please make a note in your post.
- Your Interaction with your teammates (10 points, due 9/30)
 - You must have at least **five** comments or responses in total (2 points for each).
 - You must comment on at least one post made by your teammates.
 - You must respond to at least one comment made by your teammates on your post.
 - Your comment or response must be clearly written and relevant.

This learning activity is organized by teams. You are allowed to participate in your own team only.