STAT 1361: Statistical Learning and Data Science
University of Pittsburgh

Data Science Project Proposals

The first part of the semester-long data science project will be coming up with the project itself. You should think "big-picture" about what kinds of questions you find interesting and how you could investigate them using data. These can involve any topic whatsoever, so long as there is reason to believe that there exists some dataset (either that is already compiled or that can be compiled) that can be analyzed to provide insights. Datasets chosen should not be 'simple'; they should contain at least hundreds (ideally thousands) of samples/examples and several (possibly even hundreds or thousands) of predictors.

Importantly, you must provide proof that you've actually downloaded and begun to explore each dataset. **In addition to the project proposals themselves, you should also include R files showing some brief output to prove that the data exists and is in a usable form.** These don't need to be overly detailed, but should prove that you have a good handle on what the data looks like. You may include some simple plots if you like. The `str` function in `R` is a useful tool to quickly see a summary of a data frame. Most importantly, check to see how many variables there are, how many you might want to use (i.e. are any repetitive), how much data is missing, why might that data be missing (e.g. does a missing value possibly correspond to something like 0), etc. Have a look at the applied exercises from ISLR Chapter 2 for some further ideas of the kinds of things you can do to explore the data.

You are required to propose 2 different projects. Ideally, one project should involve (primarily) regression and one should (primarily) involve classification. However, there can be exceptions to this so long as a problem normally seen as a regression problem could naturally be transformed into a classification problem or vice versa. Each project proposal should include a tentative title, a concise description of the big (overall) idea of the project, a description of the relevant data to be used, and a paragraph outlining various specific questions of interest that could be addressed. Each of these components is described below, followed by an example. Your proposals should all look very similar to the example provided here.

Title or short description: A tentative title for the project

Big Ideas: In 1-2 sentences, describe the primary or motivating goal of the project

Data Description: This needs to be concrete and involve either specific information on an existing dataset, or specific ways in which the data can be obtained. How many variables (features) are involved? What is the response? What are some of the key features of interest? How large is the sample? Is it in a format that can be nicely read into R? Is any data missing? Can the data be used directly or will it require any preprocessing (i.e. relabeling, removing/imputing missing values, etc.)? Any other relevant notes.

Questions of interest: What kinds of questions do you hope to be able to answer? Is the

problem supervised, unsupervised, regression, classification etc.

Interested Parties: Who might be interested in this kind of problem and/or data? If a person or organization hired you to take on a project like this, who might that person/organization be? **Example:**

1. **Title:** Is this mushroom safe to eat?

   Big Idea We will investigate how certain we can be that a given mushroom is edible given a variety of information about its appearance.

   Data Description: The data can be downloaded in csv format from `https://www.kaggle.com/uciml/mushroom-classification`. The data contains a total of 8124 observations on 23 features and 1 response and there are no missing values.

   - **Response:** Categorical with 2 classes: edible (e) or poisonous (p).
   - **Features:** 23 total, all categorical. We anticipate that some interesting features may be **odor** (the class of odor observed from the mushroom), **cap shape** (the shape of the cap/head of the mushroom), and **cap color** (the color of the cap/head of the mushroom)
   - **Notes:** All features are categorical, though some may be easier to treat and re-code as numeric. For example, *ring-number* is a feature which counts the number of rings and in the original data is classified as either none (n), one (o), or two (t). It may be better to relabel these as simply 0, 1, 2 and treat the feature as numeric.

   Questions of interest: The primary question of interest will be to determine how well we can predict whether the mushroom is safe to eat given the feature information provided. It may also be of interest to explore the degree to which certain individual (or small groups) of features can predict whether the mushroom is safe to eat? For example, it may be the case that we can predict whether the mushroom is safe to eat almost as well using only the odor or only the odor and shape as features. We can also explore how well predictions can be made using only the information an everyday person would be able to detect, like cone shape, odor, whether or not bruising is present, etc. as opposed to using the more advanced features like gill spacing.

   Interested Parties: This might be something of interest to something like a National Park Service. One could imagine a situation where a large number of people had gotten sick from eating poisonous mushrooms and they planned to use the results of this analysis to put together a brochure for the public on which mushrooms would be safe to eat.

Where to find data?

Prior to submitting your proposals, it is crucial that you understand the data that you intend to work with. In almost every case, this means downloading the actual dataset, loading it into `R`

and taking a good look for yourself. Only in rare cases should you propose a project in which the data will need collected and/or assembled and in these cases, very explicit instructions should be provided as to how the data will be obtained and what the final dataset will look like. If you have an idea for a project like this for which data will need obtained, you should speak to me in person prior to proposing it.

The web contains vast amounts of data. Several examples of good places to start looking for data are listed below, but you are absolutely not limited to these sites.

1. Western PA Regional Data Center: `https://data.wprdc.org/dataset`
   Hosts a variety of publicly available datasets related to activities of Western PA and Pittsburgh in particular.

2. Kaggle: `https://www.kaggle.com/datasets`
   `Kaggle.com` is a website that routinely hosts predictive competitions. Datasets that have been used in former and ongoing competitions can be downloaded and often comes with nice overviews and descriptions.

3. UCI ML Repository: `https://archive.ics.uci.edu/ml/datasets.html`
   The UCI Machine learning repository is one of the first web collections of large-scale datasets. These datasets are common choices for researchers who want to test new statistical and machine learning methodologies on 'real' datasets.

4. Enigma Public: `https://public.enigma.com/`
   The self-proclaimed "world?s broadest collection of public data." Lots of data in relatively clean form. Some data may require you to sign up for a free account before accessing.