Assignment 2: Fragile Families Challenge

Barbara Engelhardt, Princeton University out 03/01/2018; due 03/27/2018

Background

Much of this information is pulled verbatim from the Fragile Families Challenge website:

> `http://www.fragilefamilieschallenge.org/`

Please visit this and the sites mentioned in **Resources** for background and details on this project. We include a summary of that information here:

> The Fragile Families & Child Wellbeing Study is following a cohort of nearly 5,000 children born in large U.S. cities between 1998 and 2000, roughly three-quarters of whom were born to unmarried parents. We refer to unmarried parents and their children as "fragile families" to underscore that they are families and that they are at greater risk of breaking up and living in poverty than more traditional families.
>
> The core Study was originally designed to primarily address four questions of great interest to researchers and policy makers: (1) What are the conditions and capabilities of unmarried parents, especially fathers?; (2) What is the nature of the relationships between unmarried parents?; (3) How do children born into these families fare?; and (4) How do policies and environmental conditions affect families and children?
>
> The core Study consists of interviews with both mothers and fathers at birth and again when children are ages one, three, five, and nine. The parent interviews collect information on attitudes, relationships, parenting behavior, demographic characteristics, health (mental and physical), economic and employment status, neighborhood characteristics, and program participation. Additionally, in-home assessments of children and their home environments were conducted at ages three, five, and nine. The in-home interview collects information on children's cognitive and emotional development, health, and home environment. Several collaborative studies provide additional information on parents' medical, employment and incarceration histories, religion, child care and early childhood education.
>
> A fifteen-year follow-up wave includes a collection of in-home and telephone survey data from caregivers and teens.

Project definition

Your challenge is to use the Fragile Families data in an entirely new way. Given all of the data on these families from birth to year 9, and some training data from year 15, your goal in this project is to predict six key outcomes in the year 15 test data. These outcomes—**grit, GPA, material hardship, eviction, job loss, and job training**—include three binary and three continuous outcomes, and are thought to be particularly important scientifically.

Your goal in this homework project is to use a data set consisting of several thousand family samples across 12,000 features to predict the six key outcomes. Here are the steps to get started:

- 3/1: You will receive an email through Princeton's SecureSend with a link to download the data file. This data file will be encrypted (see next step). Please contact Greg Gundersen if you do not receive the file by end of day:

  `ggundersen@princeton.edu`

- 3/2: In precept on Friday, Matt Salganik and Ian Lundberg will present the Fragile Families project and data and also tell you the encryption password to access the data.

- There are some item non-responses in the birth to year 9 data. In other words, not everyone answered every question. This is a common pattern in survey data, and you will learn more about it in precept on Friday. To deal with this item non-response you can use our imputation script to fill in missing values in the data, or write your own script. For more information about missing data, you can read this blog post:

  `http://www.fragilefamilieschallenge.org/missing-data/`

- Build models to predict the six key outcomes from the other data. There are three acceptable directions to go with this:
  - Build a model to predict and analyze one of the six outcomes very well.
  - Build a model to predict and analyze one of the two types of outcomes (binary/continuous) well.
  - Build a model to predict and analyze all of the key outcomes well.

- Predict outcomes for the test set. This can be done as many times as you would like, but you can only upload up to 10 submissions per day. See **Resources** for how to submit to the leaderboard. The scoreboard will be continually updated to reflect relative performance for each of the six outcomes separately.

- Write up the project as described below (this will be the same format as the previous project).

- If you are willing to make your work open source, please upload your code, predictions, and writeup to the Fragile Families submission site using these guidelines:

`https://tinyurl.com/ffchallenge-reproducibility`.

Essential to any data analysis task is the interpretation of the results. What features were most important for prediction, and what do these features tell us about the problem? Were some reference samples more predictive of the held out sample than others? What feature selection approaches improved the results, and what methods made the results worse? The competition is very interested in identifying children that have *beaten the odds*; are there children for whom the predictions do not match the truth (even in the training samples), and what is anomalous about those children? Because of the difficulty of interrogating the actual residuals, please include in your writeup your guesses as to which features/approaches will be most informative, and also which of the key outcomes will be most difficult to predict.

Deliverables

Your deliverables for this project include:

- A five page (not including citations) summary of the project work, which should contain (as described in the Example project write up on Piazza):
  - A title, authors' names, and abstract for the project;
  - an introduction to the problem being addressed;
  - a description of the data;
  - a description of the methods developed and used, and how they were fitted using reference data;
  - a one page complete description of a specific model and associated method used for this analysis;
  - a presentation of the results of the methods applied to the test data (in whatever form that takes from the FF competition website);
  - a discussion of the results, including specific examples of single features that highlight the behavior of the models use to predict the 6 key year 15 outcomes;
  - a short summary and conclusion, including extensions that you believe would be particularly valuable based on the results;
  - a *complete* bibliography to support the methylation databases, feature selection, prediction methods, code bases, and related work that are relevant to your project.

Please use the LaTeXtemplate we have provided for you. Put your PDF write up of the project, your code, and your predictions into the Fragile Families Competition website by 5pm on the assignment due date. Name your writeup: `<author1NetID>_<author2NetID>_hw2.pdf`. Please only submit one PDF per pair of authors.

We strongly recommend *writing as you go* in the project, which means starting to write the project report as you are downloading and analyzing the data. That said, you should avoid speculative writing, and only write results once you have them.

Submission

Please put your PDF write up of the project, the Python code for the project, and a readme about how to run the code into `https://dropbox.cs.princeton.edu/COS424_S2018/Assignment2` by midnight on the assignment due date, with the following file names:

- For the write up: `<author1NetID>_<author2NetID>_hw2.pdf`
- For the code and readme: `<author1NetID>_<author2NetID>_hw2.zip`

Extensions

If you would like to extend this assignment to more interesting ground after first completing the basic deliverables for the project, you might consider the following:

- *More complex models*: there are a number of more sophisticated statistical models that might be used for this task, including, e.g., Gaussian processes [Roberts et al., 2013, Deisenroth et al., 2008], generalized linear models, or something of your own design that might exploit latent structure in the data, such as the time series, to improve predictions.
- *Confidence intervals*: Prediction may be much easier for some metrics than for others; quantifying uncertainty in prediction values is an interesting extension that would be useful for downstream tasks.
- *Predicting prediction results*: Will you enter your prediction without having an idea of how it will perform? (this is fine!) Or, will you strategically split the training set to have an internal, representative validation set to estimate whether updates to your approach will positively impact test error? You might also use this internal validation set to more carefully analyze the features and prediction uncertainty.
- *Study the residuals of these models*: Which participants were hard to predict? Which of the participants performed way better than expected and beat the odds? Is there any way to predict what children will overperform? The FF Competition website has a number of ideas for additional problems that can be addressed.
- *Use subsets of the data*: The dataset can be divided in many ways. Can you make more accurate or interpretable predictions using different subsets of the data? For example, how does using data from just the mother compare with using data from just the child?

Resources

Please spend time on the Fragile Families Competition website, which has additional details and ideas on the data:

`http://www.fragilefamilieschallenge.org/`

Submit your predictions using this website (you will need to create an account):

`https://codalab.fragilefamilieschallenge.org/competitions/20`

The Fragile Families metadata is available through both a website and a web API:

`http://browse.fragilefamiliesmetadata.org/`
`http://api.fragilefamiliesmetadata.org/`

There are both R and Python packages for interacting with the web API:

`https://github.com/fragilefamilieschallenge/ffmetadata`
`https://github.com/fragilefamilieschallenge/ffmetadata-py`

Finally, feel free to email the Fragile Families organizers with questions about the data or project that cannot be answered via Piazza:

`fragilefamilieschallenge@gmail.com`

References

Marc P Deisenroth, Jan Peters, and Carl E Rasmussen. Approximate dynamic programming with gaussian processes. In *American Control Conference, 2008*, pages 4480–4485. IEEE, 2008.

Stephen Roberts, Michael Osborne, Mark Ebden, Steven Reece, Neale Gibson, and Suzanne Aigrain. Gaussian processes for time-series modelling. *Phil. Trans. R. Soc. A*, 371(1984):20110550, 2013.