CSE8803 Projects: Big Data Analytics for Healthcare

Jimeng Sun

***Abstract*—CSE8803 Big Data Analytics for Healthcare is a graduate level course focusing on practical big data technology for health analytic applications. One big part of this course is to conduct an individual project that addresses a real-world data science problem in healthcare. The project should provide an end-to-end coverage of data science activities in addressing a real healthcare problem. The project should utilize big data systems such as Hadoop and Spark, machine learning algorithms that are covered in this class and real-world health related data. I hope that the best projects (with some additional effort) can lead to publications at the best medical informatics venues such as Journal of the American Medical Informatics Association (JAMIA), Journal of Biomedical Informatics (JBI), Journal of Medical Internet Research (JMIR), Artificial Intelligence in Medicine, IEEE Journal of Biomedical and Health Informatics (JBHI).**

This document provides the project guideline such as expectation, timeline, deliverables. We also introduce recommended project topics for selection but you are welcome to propose your own project as long as they are related to big data technology covered in this course and addressing healthcare problems.

***Index Terms*—Big data, Health analytics, Data mining, Machine learning**

I. Introduction

BIG data and healthcare applications interact closely nowadays thanks to the advancement in electronic data capturing technology such as electronic health records, on-body sensors and genome sequencing. This course is about learning practical skills on big data systems, scalable machine learning algorithms and their applications to healthcare. Through (painful) homework exercises, all the students should have by now learned big data systems and acquired sufficient knowledge about healthcare data. We believe you are ready to take on the next level of challenges as a data scientist in healthcare. That is, you are going to propose, execute and report an awesome data science project. The final results of this project includes **1) a publishable report and 2) a convincing presentation, and 3) reusable software and sufficient documentation from your project.**

Next we will cover the project life cycle, timeline, deliverables, grading scheme and project topics.

II. Project Life Cycle

As a data scientist working on a real-world project, you have to be able to conduct all aspects of the big data project independently in a timely manner. In particular, here are some tasks that a data scientist will have to conduct in a big data project: project initiation, project execution and project report.

J. Sun is with School of Computational Science and Engineering Georgia Institute of Technology, Atlanta, GA, 30332 USA e-mail: (see http://www.sunlab.org).

A. Project initiation

As a data scientist, projects are not always there for you to work on. You have to create them and convince your boss (e.g., your CEO) to fund that. Before your project is officially launched, you have to conduct many steps to make that happen. Here are the checklist of things that you should do during the project initiation.

1) Identify and motivate the problems that you want to address in your project.
2) Conduct literature search to understand the state of arts and the gap for solving the problem.
3) Formulate the data science problem in details (e.g., classification vs. predictive modeling vs. clustering problem).
4) Identify clearly the success metric that you would like to use (e.g., AUC, accuracy, recall, speedup in running time).
5) Setup the analytic infrastructure for your project (including both hardware and software environment, e.g., AWS or local clusters with Spark, python and all necessary packages).
6) Discover the key data that will be used in your project and make sure an efficient path for obtaining the dataset. This is a crucial step and can be quite time-consuming, so do it on the first day and never stops until the project completion.
7) Generate initial statistics over the raw data to make sure the data quality is good enough and the key assumption about the data are met.
8) Identify the high-level technical approaches for the project (e.g., what algorithms to use or pipelines to use).
9) Prepare a timeline and milestones of deliverables for the entire project.

All the above steps in project initiation should be demonstrated in your proposal.

B. Project execution

Once your project is approved, you should quickly work on getting results and iterate with your sponsors on the progress. Iteration is the key. The first iteration should be fast and positive otherwise you are at risk losing momentum from the sponsors/project owners (e.g., your boss, clinical experts, your partners from another organization). This successful execution

will lead to long-term sustainability of your team and will greatly improve your reputation in the organization, so please focus on that.

1) Gather data that will be used in your project if you haven't already.
2) Design the study (e.g., define cohort, target and features; carefully split data into training, validation to avoid overfitting)
3) Clean and process the data.
4) Develop and implement the modeling pipeline.
5) Evaluate the model candidates on the performance metrics.
6) Interpret the results from your model (e.g., show predictive features, compare to literature in terms of finding, present as cool visualization).

All the steps in project execution should be done by the paper draft due day and iterate at least another time by the final due day.

C. Project report

Finally, you are close to the end of the project. You need to summarize what you have done and learned throughout the project. This will be a comprehensive, concise and well-written report as the foundation for future projects. This can lead to publications and other external communication. You will probably need to give a presentation to your sponsors. So do the best you can in written report and presentation. Bad delivery at this step can overshadow all the great work your team have put in throughout the project, so do spend sufficient time to prepare a slick presentation and write a comprehensive report.

- Your report should consists of the following sections.
  1) Title and abstract
  2) Introduction and background
  3) Problem formulation
  4) Approach and implementation
  5) Experimental evaluation
  6) Conclusion
- Prepare a presentation deck and deliver a convincing and informative presentation.
- Clean up and package your code, and document the necessary steps for future usages by others.

Please use the above process to guide your own project for this semester and possibly your future data science career.

III. Logistics

Next we summarize the timeline and deliverables for your project in this semester.

A. Timeline

Due daterrrr	Task description
Oct 23	Project proposal submission
Oct 30	Peer review bidding
Nov 20	Project draft
Nov 27	Peer feedback for draft
Dec 4	Final paper and presentation
Dec 11	Peer feedback for final project

Note that for peer review bidding, you need to login into easychair system for CSE8803BDH Workshop Fall 2016 and bid on 5 papers as willing to review. Eventually, 3 papers will be assigned to you for review. You will receive an email to be a PC member from easychair, and make sure you login and bid papers.

B. Deliverables

1) Project Proposal:

- 1-page write-up (word and latex templates will be provided)
- Guide:
  - Explain about the problem/topic you choose and how to solve it
  - It is recommended to try to cover as many aspects as described in project initiation if it is possible.
  - Conduct literature search and cite at least 4 papers that are relevant to the project.

2) Project draft:

- up to 4-page write-up + 1 page reference
- Guide
  - Make sure your write-up cover all aspects described in project execution.
  - Conduct literature search and cite at least 8 papers or more that are relevant to the project.

3) Peer feedback on others' paper draft: You need to assess other's work based on the following criteria:

- Presentation quality
- Importance of the problem
- Comprehensive literature review
- Feasible and meaningful approach
- Clearly identified evaluation metric

4) Final report:

- up to 5-page write-up + 1 page reference (the same template as project draft).
- 5-min presentation (youtube or audio attached slides) + slides.
- software implementation and documentation

C. Grading scheme

Here are the grading guideline for your project and peer participation.

- Project 45%
  - 5% proposal
  - 10% paper draft
  - 12% final presentation
  - 18% final paper (including Kaggle result)
- Peer feedback (how you reviewed others papers) 6%
  - 2% draft paper
  - 4% final paper and presentation

IV. Project Topics

We introduce several project topics for your consideration but you can also propose your own project outside this scope as long as your project uses big data tools (e.g., Hadoop and Spark) and is about healthcare applications.

A. Treatment Recommendation and Refractory Patient Management in Epilepsy

Mentor: Sungtae An ⟨stan84@gatech.edu⟩

For these projects, you will develop predictive model for helping epilepsy patients. Two types of models should be considered 1) predicting refractory epilepsy patients and 2) predicting the treatment effectiveness using big data tools (Hadoop and Spark). We suggest that you target one of the predictive models in your project. However, you can also propose your own project (upon approval) utilizing our large epilepsy dataset with big data analytics tools.

1) Predictive Modeling of Refractory Epilepsy Patients: Drug-resistant epilepsy is a major clinical and societal problem for one in three epilepsy patients. We call this drug-resistant epilepsy population as refractory epilepsy patients. More specifically, we can define the refractory epilepsy population as epilepsy patients who failed with 3 distinct anti-epilepsy drugs (AED). In this project, you will work on building a predictive model to determine whether a patient will likely be a refractory epilepsy patient in advance, especially at their first AED failure time. Please refer to this Description for more details. This project includes Kaggle in Class competition among students who choose this topic.

2) Treatment Recommendation for Epilepsy Patients: In this project, you will develop a model to recommend optimal epileptic treatment regimens to patients. For this purpose of the project, you can use any kind of machine learning techniques. For example, you can build a predictive model for each regimen, which consists of a single AED or multiple AEDs, separately. Please refer to this Description for more detail information. We show an example approach in the description, but it is not restricted to apply any method you want to use.

- **Resources**: Description for refractory epilepsy patient prediction, Description for epilepsy treatment recommendation. Here are some related papers to get you started with the clinical challenges in epilepsy treatment [4], [6], [2], [7]. You are welcome to discuss the ideas with the mentor.
- **Dataset**: It will be provided through the secure environment.
- **Metrics**: AUC, sensitivity, specificity, scalability, etc.
- **Challenges**: Exact cohort construction, advanced feature engineering and processing big data for those steps.

B. Septic shock prediction

Mentor: Ruiming Lu ⟨rlu39@gatech.edu⟩

Sepsis is a leading cause of death in the United States, with mortality highest among patients who develop septic shock. Early aggressive treatment decreases morbidity and mortality. While general-purpose illness severity scoring systems are useful for predicting general deterioration or mortality, they typically cannot distinguish with high sensitivity and specificity which patients are at highest risk of developing specific acute condition.

Using supervised learning, a machine learning methodology, and the MIMIC (Multiparameter Intelligent Monitoring in Intensive Care)-II Clinical Database (40), Henry et al. [3] trained a predictive model based on a targeted real-time early warning score (TREWScore) that identifies those patients at high risk of developing septic shock in the future. With a median lead time of over 24 hours, this scoring algorithm may allow clinicians enough time to intervene before the patients suffer the most damaging effects of sepsis.

The goal of the project is to repeat and improve the predictive model from [3] using MIMIC-III database [5] with big data tools (Hadoop and Spark). The analytic process should follow the predictive modeling pipeline covered in the lecture. The detailed steps should be presented including prediction target, cohort construction, feature construction, feature selection, predictive model and performance evaluation. The key phases are contructing the features and building the predictve model.

The following are some guidelines for this project, but it is OK to deviate from the guidelines as long as you clearly state what you did and why it makes sense. For cohort construction, you can use ICD codes, criteria given in the paper, or other meaningful method. There is no date associated with the diagnostic cod. Therefore, you have to rely on looking at when the set of criteria given in the paper are satisfied to determine the time of onset of septic shock. Any other method are also welcome. Henry et al. [3] obtained With a median lead time of over 24 hours to predict, feel free to start off using that as the length of your prediction window and see how your model performs compared to what the paper obtained. The performance should be validated by cross-validation or a separate hold-off set. It is also expected that there are some scalability results reported, e.g., runtime vs no. of patients, runtime vs no. of cores, since you are using big data tools in this project.

This project includes a Kaggle in Class competition among students who choose this topic.

- **Resources**: Supplemental material for the paper.
- **Dataset**: MIMIC III.
- **Metrics**: AUC, sensitivity, specificity, scalability, etc.
- **Challenges**: Feature construction and implementation using big data

C. Mortality Prediction in ICU

Mentor: Nisheeth Bandaru ⟨nisheeth@gatech.edu⟩

"Accurate knowledge of a patients disease state and trajectory is critical in a clinical setting. Modern electronic healthcare records contain an increasingly large amount of data, and the ability to automatically identify the factors that influence patient outcomes stand to greatly improve the efficiency and quality of care.

Ghassemi et al. [1] examined the use of latent variable models (viz. Latent Dirichlet Allocation) to decompose free-text hospital notes into meaningful features, and the predictive power of these features for patient mortality. This work considered three prediction regimes: (1) baseline prediction, (2) dynamic (time-varying) outcome prediction, and (3) retrospective outcome prediction. In each, our prediction task differs from the familiar time-varying situation whereby data accumulates; since fewer patients have long ICU stays, as we

move forward in time fewer patients are available and the prediction task becomes increasingly difficult."

The goal of this project is to repeat and improve all or part of the study[1] using the newer MIMIC-III data[5] implemented with big data tools (e.g., Hadoop and Spark). You must present detailed steps such as the prediction target, feature selection, feature construction, predictive model and performance evaluation.
You may inititally start with a small subset of data as you develop your model locally. However, after fine-tuning it, your final paper must be based on results from the entire data. You may also want to narrow your focus on implementing a subset of the models discussed in the paper. The model must be evaluated by cross-validation and its performance on a hold-off test set. This project includes a Kaggle in Class competition among students who choose this topic.

- **Resources**: Paper, Presentation, Video.
- **Dataset**: MIMIC III.
- **Key Deliverable**: Models built with MIMIC-III data with similar or better performance.
- **Metrics**: AUC, sensitivity, specificity, scalability, etc.
- **Challenges**: Feature construction and implementation using big data tools.

D. OHDSI Analytics with Big Data

Mentor: Adrian Chang ⟨adrian.chang@gatech.edu⟩, Peter Schneider ⟨peteryschneider@gatech.edu⟩

The Observational Health Data Science and Informatics (http://www.ohdsi.org/analytic-tools) is a community that focus on developing open source health care analytic tools. Unfortunately, most of those tools are built around traditional technologies are not designed to handle the large volume of data that modern healthcare creates.

For this project, you can either develop a big data version of an existing OHSDI tool or create a new one based on the big data tools (Hadoop, Spark) that we have learned so far in this class. For example, if you were to choose ACHILLES to rewrite, part of your project would be to rewrite all of their SQL data processing in Spark. While developing the tool, you must also consider the data format (OMOP) and the way OHDSI tools store their data and see if either can be improved upon in addition to the actual data processing. By the end of the project, your tool should be able to be reused by other people and be polished enough that it could be published as an OHDSI tool. You are welcome to conduct your project leveraging the data such as MIMIC III and CMS synpuf.

If you choose this project, your project will primarily be evaluated on the following:

1) Reusability
   Does your tool produce consistent results that require little effort to setup and reproduce?
2) Scalability
   What is the limit of data that your tool can handle? Is it much better than an equivalent tool?
3) Accuracy
   Does your tool produce comparable if not better results than the original tool or an equivalent tool? Did this occur because you chose a different algorithm to produce your results or because of something else?
4) Ingenunity
   Is your tool taking a same approach to solve the same problem? Or did you invite something new?

- **Key Deliverable**: Reusable big data healthcare software transplanted from OHDSI.
- **Input Data**: OMOP compliant database (e.g,. MIMIC III and CMS synpuf).
- **Metrics**: Scalability, accuracy, reusability.
- **Challenges**: Design and implementation of OHDSI tools and familiarity with the OMOP data format.

E. Other projects

You are welcome to propose your own projects as long as 1) they are health analytic projects and 2) they use big data tools covered in this class (Hadoop, Spark). Note that we will not provide much support on those projects.

V. Conclusion

Best of luck on your project and data science rocks!

Acknowledgment

Thanks all the TAs for their time and effort in creating the course material together. Thank all the students for their dedication and feedback.

References

[1] M. Ghassemi, T. Naumann, F. Doshi-Velez, N. Brimmer, R. Joshi, A. Rumshisky, and P. Szolovits. Unfolding Physiological State: Mortality Modelling in Intensive Care Units. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 75–84, New York, NY, USA, 2014. ACM.

[2] I. Gilioli, A. Vignoli, E. Visani, M. Casazza, L. Canafoglia, V. Chiesa, E. Gardella, F. La Briola, F. Panzica, G. Avanzini, M. P. Canevini, S. Franceschetti, and S. Binelli. Focal epilepsies in adult patients attending two epilepsy centers: classification of drug-resistance, assessment of risk factors, and usefulness of "new" antiepileptic drugs. *Epilepsia*, 53(4):733–740, Apr. 2012.

[3] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, Aug. 2015.

[4] P. Kwan and M. J. Brodie. Early identification of refractory epilepsy. *The New England Journal of Medicine*, 342(5):314–319, Feb. 2000.

[5] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. Multiparameter Intelligent Monitoring in Intensive Care II: a public-access intensive care unit database. *Critical Care Medicine*, 39(5):952–960, May 2011.

[6] D. Schmidt and W. Lscher. Drug resistance in epilepsy: putative neurobiologic and clinical mechanisms. *Epilepsia*, 46(6):858–877, June 2005.

[7] A. Voll, L. Hernndez-Ronquillo, S. Buckley, and J. F. Tllez-Zenteno. Predicting drug resistance in adult patients with generalized epilepsy: A case-control study. *Epilepsy & Behavior: E&B*, 53:126–130, Dec. 2015.