MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

FSRP: Implementing Modern Machine Learning Models for an Automated Data Science Pipeline

Summary: As a part of the DARPA D3M program, Professors Jason Corso and Laura Balzano are leading an effort to implement modern machine learning models and techniques to become modules in an automated data science pipeline. Concretely, there are ten datasets and related machine learning tasks for which we will be implementing possible solutions in Python. The types of methods include such as matrix completion, sparse coding, subspace clustering, multimodal dictionary learning, and novel deep learning architectures. Other teams in the program will be building automated pipelines to test a combination of ML techniques for each dataset/task and select the best in terms of some given performance metric. At an integration meeting at the end of the summer, all teams will work together to integrate an overall system for best performance. This project will give you experience with modern ML methods as well as experience integrating into an exciting system along with a large team of data scientists.

Items	Description	Notes
Affiliation *	Electrical Engineering and Computer Science, College of Engineering	
Mentor *	Jason Corso, Associate Professor, jjcorso@umich.edu, http://web.eecs.umich.edu/~jjcorso/ Laura Balzano, Assistant Professor, girasole@umich.edu, http://web.eecs.umich.edu/~girasole	
Logistics *	5 months – 1 year Team Project for 3-5 students. Start date: May 1 or earlier, Finish date: Sept 30, can be extended if there is mutual interest.	
Scope *	Data Science techniques are varied and technical, and typically a domain expert who works with data must also have extensive experience with statistics and machine learning. The DARPA D3M program hopes to somewhat automate the data science process in three parts: 1. build basic modules (starting with the ML modules available in SKLearn Python library), 2. create systems to stack, organize, and test combinations of those ML modules, and 3. interact with a human domain expert to narrow down potential models and/or identify the real interesting task at hand. Our role in this project is part 1, building basic ML modules for use in the pipeline.	
Challenge *	Machine learning on varied data, with missing data, etc, and implementing techniques to be broadly applicable.	
Data *	There are 10 data science problems, ranging to predicting baseball hall-of-famers from their stats to classifying sounds from an Urban audio recording to predicting miles-per-gallon from other automotive characteristics. The data include tabular, text, audio, and image.	
Approach	We will implement specific techniques such as matrix completion, sparse coding, multimodal dictionary learning, and classification, and examine their performance on these datasets.	
Students	Students on this project will desire experience implementing modern ML methods on a variety of interesting datasets. They will have programmed in Python before.	
Expectations *	We expect each student to implement 3-4 methods in Python, document the code in detail, and demonstrate their performance on the relevant datasets in our collection. We expect the students to participate in the integration meeting. Finally, we welcome students who wish to develop novel approaches or implementations; in that case we will expect fewer implementations but a more significant report detailing the ideas behind the new method.	
Products *	Software and documentation.	
Format *	The mentors will meet with the students every other week as needed, and the students will also work together on the project.	
Platform	The code will be in Python and will integrate with the DARPA D3M schema.	
Funding	$18 per hour	