MACHINE LEARN YOUR WAY TO MARCH MADNESS GLORY!

Teaching Biochemists and Biophysicists Machine Learning

Introduction/Background

IMPORTANCE OF MACHINE LEARNING TO BIOCHEMISTRY AND BIOPHYSICS

Biochemistry and biophysics are two fields that are ripe with many exciting breakthroughs. Machine learning, a type of artificial intelligence where computer programs adapt to new data, is used by biochemists and biophysicists to do things like generate new strands of DNA. Our client, a professor in the department of Biochemistry and Biophysics at OSU, recognized there was a need for his budding scientists to understand machine learning so they could be better prepared for their careers.

A NEED FOR A MACHINE LEARNING INSTRUCTIONAL TOOL

Our client noticed that the Biochemistry and Biophysics curriculum at OSU did not encourage undergraduate students to learn machine learning. Even if machine learning classes were to become a cornerstone of their coursework, the content would be difficult for people without a Computer Science background. To make matters worse, teaching machine learning to these students through its application to biochemistry and biophysics is particularly challenging, since machine learned models of DNA can be hard to interpret.

WHAT WE WERE COMMISSIONED TO DO

We were enlisted to produce an online instructional module where these students could grasp machine learning fundamentals in a fun and clear manner. Our client wanted us to develop this module so that students could generate machine learned NCAA March Madness brackets. Since a fundamental aspect of learning machine learning is recognizing how the inclusion or exclusion of data influences resulting models, this module would satisfy the need by producing models (brackets) that were distinguishable from each other based on the college basketball statistics a user chose to train their model on. In addition to this clarity, the module would also be fun to use, because rather than focus machine learning on biochemistry and biophysics, it would instead concentrate on men's college basketball.

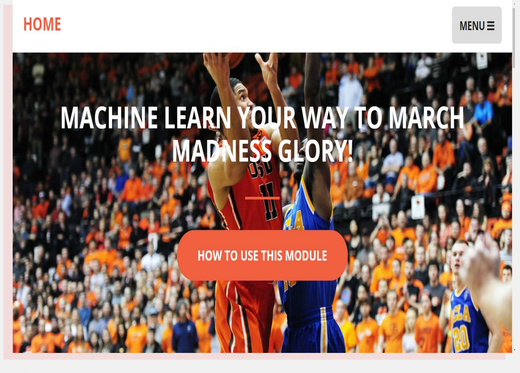

Fig. 1: Home page. The "Menu" button allows the user to navigate through the website.

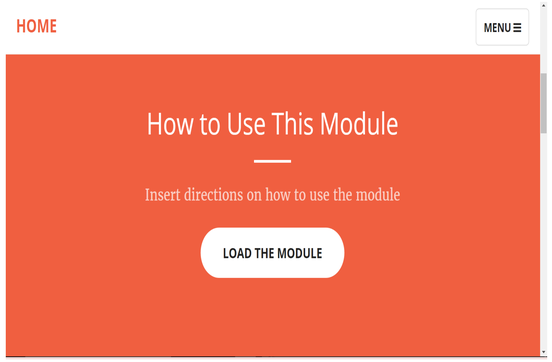

Fig. 2: Instructional page that informs users on how to use the machine learning module.

PROJECT INFORMATION

Class: CS Senior Capstone, 2016-2017

Developers:

- **Alex Hoffer** (hoffera@oregonstate.edu)
- **Jacob Smith** (smitjaco@oregonstate.edu)
- **Chongxian Chen** (chencho@oregonstate.edu)

Client: Dr. Victor Hsu, Oregon State University, Department of Biochemistry and Biophysics

PROJECT DESCRIPTION

To implement the module, we needed to complete the following five steps:

1. Develop a Graphical User Interface (GUI)
2. Aggregate/select college basketball statistics
3. Feed statistics to machine learner
4. Train a model
5. Generate bracket of model

The following headings are technical descriptions of the five steps:

GUI

Alex used HTML, CSS, and JavaScript to produce the GUI for our web page. HTML was used to split the page into logical sections such as Home (found in Figure 1), Instructions (found in Figure 2), Module (found in Figure 4), Purpose, and About (found in Figure 3). We utilized CSS to make these sections look clean and usable. Finally, JavaScript was used to enhance the user experience by making the page interactive, such as turning certain buttons different colors upon clicking in order to notify the user of the action they had just performed.

AGGREGATE/SELECT STATISTICS

College basketball statistics from 1985 to the current season were gathered from the websites Kaggle.com, ESPN.com, and NCAA.com in the CSV file extension and stored using PHP MyAdmin. This tool was selected because Jacob was familiar with it and found it usable. Since the regular season didn't conclude until March, Jacob manually updated the database to reflect the current standings frequently until the final game was played.

FEED STATISTICS TO MACHINE LEARNER

Using Python, Chongxian gathered the statistics the user selected and fed them into his machine learning module.

TRAIN A MODEL BASED ON STATISTICS

Keeping with the Python theme, Chongxian used Scikit-Learn, an open source Python library, to generate a machine learned model based on the statistics chosen by the user.

GENERATE BRACKET OF RESULTS

The machine learned model was then transferred into bracket form by Jake (incomplete).

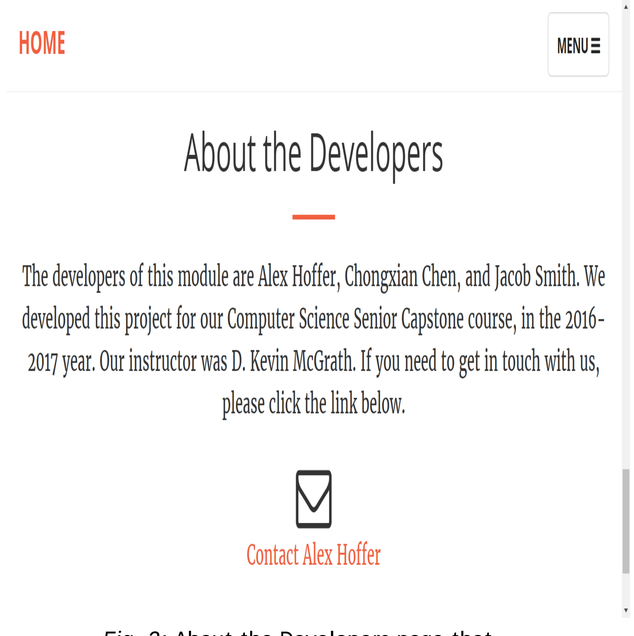

Fig. 3: About the Developers page that includes our information.

CONCLUSION

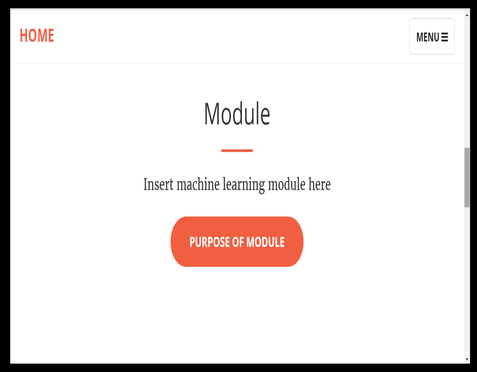

Fig. 4: Module page that hosts the machine learning module.

FEATURES PROVIDED BY MODULE:

- A Graphical User Interface (GUI) including a home page, an instructions page, an about the developers page, and a statement of purpose.
- The ability to select from a set of college basketball statistics.
- A machine learned bracket that corresponds to the specific statistics the user requested the model to be trained on.

As of mid-March 2017, we can conclude that the project is incomplete. The machine learning module has only been partially implemented by Chongxian. This has had a number of different effects: the user is not able to generate predictions, and as such, the other developers have not been able to work on converting the generated models into brackets. Once the final pieces are put into place by Chongxian, Jake can work on pretty printing the bracket, which Alex will help with. Future machine learning modules similar to this one will want to have a better way for users to select college basketball statistics. Additionally, the developers of such modules may wish to have prior machine learning experience, a more proper mode of communication, and a more rigid implementation schedule in order to allow for time at the end to polish each component of the module separately.

Oregon State UNIVERSITY