Senior Design 2016

Problem Statement

By Alex Hoffer, Jake Smith, and Chen Chongxian

Abstract

With the growing dependency of industry and research work in biochemistry on computing, it becomes increasingly important for students studying biochemistry to be well-versed in state-of-the-art techniques like machine learning. Machine learning is a practice that enables biochemists to do remarkable things like discover new correlations between DNA and protein sequences. Since many of these students are burgeoning researchers or future industry professionals themselves, knowledge of machine learning is a necessity. The only problem is that these students are trained in biochemistry, not computer science. Therefore, there is a need to teach students of biochemistry about machine learning in a simple, clear, and entertaining way. This project aims at providing such an instructional tool to the students of Biochemistry and Biophysics at Oregon State University. This instructional tool will be a module hosted online for these students to learn machine learning concepts through experimentation with datasets and training related to the NCAA Mens Division I Basketball Tournament, also known as March Madness.

1 Problem Definition

Machine learning is a useful tool for biochemical research and industry alike, but a problem is that computing skills of that kind are not emphasized in most Biochemistry and Biophysics curriculum. Additionally, grasping machine learning through its application in Biochemistry and Biophysics is a particularly difficult way of learning it. This is because biochemical results that are generated by machine learning tend to be unclear and open to interpretation, making it hard for machine learning neophytes to see the effects of including or excluding specific types of data. Therefore, students who study this subject need to have the opportunity to learn machine learning for their future jobs in this field, whether it be in industry or research. They need for this learning process to be reasonably clear because they dont have a background in computer science and machine learning approaches can be esoteric and confusing.

2 Proposed Solution

Our client is a researcher of Biochemistry and Biophysics who is motivated to make it easier for his students to understand machine learning. We will make this possible by producing an online module which will enable students of Biochemistry and Biophysics to generate their own NCAA Men's Division I Basketball Tournament brackets using machine learning. The students will select statistical categories from an expansive library we will provide to train their models with, which will allow them to predict which teams will advance in the tournament. The NCAA Mens Division I Basketball Tournament consists of 68 college basketball teams that clinch berths based on winning their respective conferences in the regular season or are selected by the NCAA committee. They are then assigned to one of four regions (South, East, West, Midwest) based on the location of the college. Teams are seeded based on their win percentage in the regular season. So, for example, a first seeded team will play the sixteenth seeded team in their region, a second seeded team will play the fifteenth seeded, and so on. Once a team wins each of their games within their region, they enter into the Final Four, where they play a corresponding regions winningest team. Finally, the two teams that emerge from these games battle each other and the winner of the final game is deemed the victor of the tournament.

It is our client's belief that the process of teams battling their way through the bracket provides a compelling opportunity to learn machine learning. Machine learning can be implemented so that a person selects which statistics they believe are useful in predicting whether a team will win their next game. These statistics are widely available on websites such as the NCAA.com and can be

compiled into massive sheets of data. As with all machine learning, the question becomes which statistics are useful in prediction and which are useless or perhaps even counterproductive. In this way, Biochemistry and Biophysics students can easily see this fact by choosing seemingly obscure statistics and clearly impactful statistics alike to train data with and see which categories are valuable to producing a good bracket and which aren't. Rather than learning machine learning through interpreting unclear biochemical data, these students will be given obvious results (a win or a loss) based on the data they choose to train their model with. Our project will consist of a user-friendly server where students can choose which basketball statistics to train data with and which statistics to avoid. These models will use past NCAA Tournament results (using specific seasons' stats and data as the training set) to assess predictive models. Therefore, this module can generate models throughout the year once it's developed. Models will be especially interesting in March, when the tournament begins, and their efforts will be easier to visualize. This solution will provide an easy interface by which these students can learn basic machine learning concepts based on predictions which are binary in nature (i.e. a team can either win or lose any particular game).

3 Performance Metrics

Our project's success will be evaluated based on whether there is a fully functional learning module hosted on a server. Functionality will be measured by the program's ability to generate a user's predicted bracket which accurately reflects the statistical categories the user chose to train data on. The client has allowed us to determine the appropriate breadth of data available for users to choose from, and has recommended certain websites to mine from. One aspect of our success, then, is having a large mixture of categories by which the user can choose from. This is necessary because the success of the instructional tool is to the extent by which the student can understand machine learning concepts, and this knowledge will be solidified in seeing how brackets change based on which sets of data they use for training. Failure will consist of a) the server does not permit users to select data categories b) the server does not generate a bracket c) the bracket generated by the server doesn't reflect the statistical categories the user selected. Therefore, we will be testing the module by selecting a representative number of statistical categories and generating brackets of our own to ensure that the brackets reflect which categories were chosen to train data with. Another component of our testing will involve whether students can properly access the module remotely. With the success of these tests, we have a guarantee that the module effectively solves our defined problem. It is important to make the distinction that the success of our module is NOT concerned with teaching these students how to apply machine learning to Biochemistry and Biophysics, but simply teaching them machine learning concepts

which they can potentially apply to these subjects given outside instruction that we will not be providing. One possibility for our presentation at the Engineering Expo is our own personal brackets generated by our module, preferably ones we have successfully been able to optimize through careful selection of statistical categories and frequent trial-and-error generation of brackets.

4 Terms and Agreement