**Implementing Logistic Regression for Protein Contact Map prediction**

**Kareith Dyce**+1**, James An**+1**, and Xavier Verdad** +1

1Auburn University, COMP, Auburn, 36849, USA \* kzd0030@auburn.edu

+these authors contributed equally to this work

**ABSTRACT**

In this project we will be using Machine Learning to predict a Protein Contact Map (MAP) This is a binary 2d list like structure that holds a 1 when the distance between two amino acids is less than or equal to a certain distance and a zero otherwise. Using Gradient Ascent, a Logistic Regression algorithm, we will be training a model so that we can be accurately predicting these PCMs.

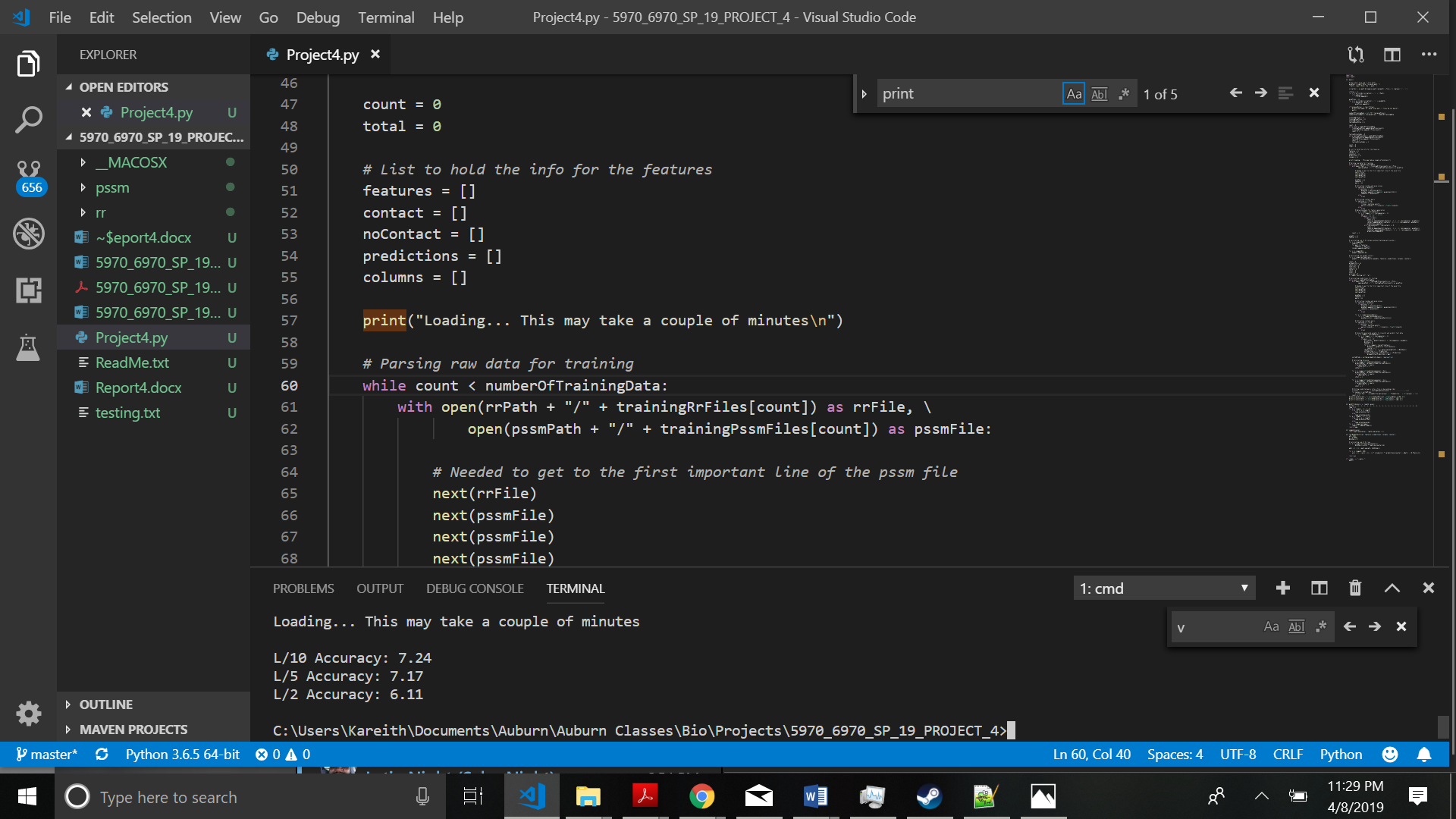
**1. Introduction**

The reason the PCM are important are that they convey strong information about the protein 3D structure, are translation and rotation invariant, and more compact and thus is more suited for learning problems.1With a PCM you are dealing with simple a binary 2d matrix. To do this we will be using machine learning technique called Gradient Ascent. This machine learning algorithm takes advantage of many logarithmic identities hence the name Logistic Regression to make a curve to the predict the binary nature of the problem.

**2. Methods**

To generate the PCM we would first need to build a model using known proteins. To do this we generate 200 features using the related pssm file and two sliding windows so that where i is the center of the first window an j is the center of the second one. For the sliding windows the pssm line relating to the selected amino acid was added as well as the lines relating to the two acids before and after. If there is not a corresponding acid, then 20 values of -1 are added to the feature. Since we have 2 sliding windows that each use a sliding window of 5 values with each having 20 features, we are able to get the 200 features as desired (2 \* 5 \* 20 = 200). Next, the add the features to four lists. One with all legal pairs, another with whether they are contacts or not and two for holding the contacts’ features and noncontact’s respectively. We need to find P(Y = 1|X,W) and P( Y = 0 | X,W), however since P(Y = 0|X,W) = 1 – P(Y = 1| X,W) we really only need to find P(Y = 1 | X). Using Baye’s Theorem and rearranging gives us . Where , and is the current feature being feed into the model. To train the model known data needs to be added to the model so that the averages of the ith features for both Y’s ( and ) as well as the variance( )for the ith feature can be calculated to predict the PCM. This is then used in the weight formula . to set the weight. To model the data You get the weight vector that was generated from the training data multiply it by the features in a residue pair to get the probability of a contact.

**3. Results**



**Figure 1.** Accuracy

**4. Discussion**

As you can see above the results that say that our accuracy is ~7% for all tests. This means that the gradient ascent was extremely inefficient in making the PCMs and guessing would have much better. The lowest accuracy was for the L/2 test while the highest was for the L/10 test.

**5. References**

1. Bartoli, Lisa et al. *Folding.Biofold.Org*, 2019, http://folding.biofold.org/pages/documents/papers/Bartoli\_MethMolBiol2008.pdf. Accessed 9 Apr 2019.