The Report of Programming Task

The aim of the task was to develop a machine-learning program, which can predict the response of HIV-infected patients to therapy. The nucleotide sequences of the Protease (PR) and Reverse Transcriptase (RT), the viral load and CD4 count of 1000 patients were provided as the training data. The accuracy of the predictions was measured using a dataset of 692 patients.

Several methods were tried. The method which is described below takes into account three of the four features provided for each patient and hence, I am more confident in its predictions. First, the distribution of responders and non-responders was examined. The ratio of the responders to non-responders in the training data was $\frac{206}{794}$, whereas this ratio in the test data was $\frac{346}{346}$. To make the population of patients in the training data similar to those in the test data, the responders' data was added thrice to the original data. This made the ratio of responders to non-responders $\frac{824}{794}$, which is roughly equal to 1.

The plot of CD4 count versus viral load suggested that viral load could be a good measure of patient's improvement. To find the threshold level of viral load that could predict patient's response to therapy, a range of different viral loads from 3 to 8 at 0.1 increments were examined. The viral load threshold level of 4.3 was found. However, this threshold level failed to make accurate predictions for viral loads between 4.2 and 5.4. As a result, the data was divided into three groups.

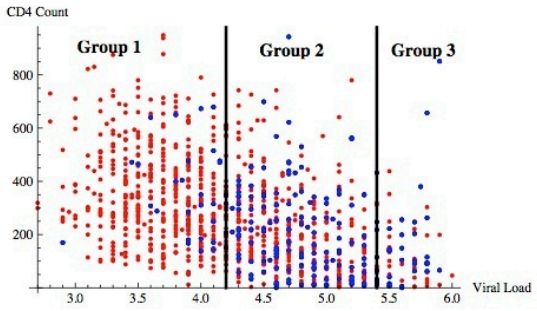

Figure 1: CD4 count versus viral load of the training data set. The figure shows the three groups of patients, which were treated separately.

Patients in the first group, with viral load less than 4.2 (and hence less than viral load threshold level) were grouped as non-responders. Those in group three whose viral load was greater than 5.4 (and hence, greater than the threshold level) were grouped as responders. For the second group, with viral load between 4.2 and 5.4, CD4 count and PR sequences were also taken into account.

To procure quantified measure of the sequences, the training data was divided into two groups: responders and non-responders. In each group, the PR sequence of each patient was aligned with the rest of the sequences in that group and the regions of similarity were identified. The sequences were aligned using the global Needleman-Wunsch alignment technique, which is suitable for sequences of the same size, such as the given PR sequences [1]. Hence, two lists were generated, which contained the common regions in the PR sequences of responders and non-responders. The intersection of these lists, which contained the regions in PR sequences found in both responders and non-responders was calculated and then eliminated from both lists. This resulted in two disjoint lists that only contained the common residues in responders and non-responders. For patients in the second group, it was calculated how many of the responders' and non-responders' common residues were found in their PR sequences. The ratio of these numbers was then calculated and plotted in a histogram. The histogram showed that the ratio peaked at different values for responders and non-responders. Consequently, it was used as the quantified attribute of the PR sequences.

Nargess Khalilgharibi

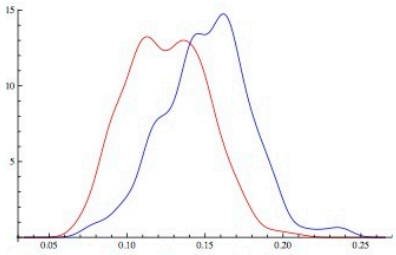

Figure 2: Histogram of the ratio of the number of responders' and non-responders' common residues found in the PR sequences of the second group.

Having 3 numeric features (VL, CD4 count and the ratio of responders' and non-responders' common residues) enabled the use of logistic regression to classify the patients in the second group. First, the two parameter logistic function was calculated using VL and CD4 count. Since the weight of the CD4 count was very small compared to the weight of the VL, changes in CD4 count did not make significant changes to the probabilities calculated by the logistic function. In order to make the weight of the CD4 count of the same order of magnitude of the weight of the viral load, different functions such as the root of the CD4 count were examined. The new logistic functions were calculated and used to predict the response of the training data. The predictions were then compared to the given response of the training data in order to calculate the accuracy of the function used. It was found that logarithm of the CD4 count to base 10 gave the most accurate predictions. Therefore, in the succeeding attempts, the logarithm of CD4 count was used to calculate the logistic function.

Since information about PR sequence of 80 patients was not provided in the training data, these patients were omitted from the training data. Also, some patients had zero CD4 count which prevented Mathematica from calculating the logarithm of it. These patients were also omitted from the training data. The ratio of the responders to non-responders in the remaining data was then $\frac{185}{733}$. Therefore, it was necessary to resample the remaining data in order to match it with the test data. For this purpose, 548 responders were added using random sampling method. A set of input, constituted of viral load, logarithm of CD4 count and ratio of the number of responders' and non-responders' common residues was fed to the LogitModelFit function in Mathematica and the logistic function with three parameters was obtained. Since there were two classes in the model, i.e. responders and non-responders, the decision boundary was set to 0.5 [2]. This meant that when the prediction probability obtained from the logistic function was less than 0.5, the patient would not respond to the therapy, whereas if it was greater than 0.5, the patient would. Having obtained the logistic function, the logarithm of CD4 count and ratio of the number of responders' and non-responders' common residues were calculated for the second group of test data and along with the viral load, were fed to the logistic function and the response to the therapy was predicted.

In conclusion, both the training and test data were divided into three groups. For the first and third group, viral load was used to predict patient's improvement. For the second group, the logistic regression was used for making predictions. The attributes taken into account to calculate the logistic function were viral load, CD4 count and PR sequence. The resulting model predicts 457 patients in the test data to be responders and 235 patients to be non-responders. It scored 69.08 on the public leaderboard. The best score of 69.56 was obtained by just looking at the viral load. As the model takes into account more features of the patients, it is believed that it will score better on the private leaderboard than the other model.

References:

1. Mount DW. Bioinformatics: sequence and genome analysis. 2nd Edition. New York: Cold Spring Harbor Laboratory Press; 2004
2. Witten IH, Hall MA. Data mining: practical machine learning tools and techniques. 3rd Edition. London: Morgan Kaufmann; 2011

Nargess Khalilgharibi