CSCI 481 Machine Learning
Assignment 5

Instructor: Murat Dundar

Due date: 5/6/2021 11:59pm

Please type in your assignment and submit it online through Canvas. No hard copies will be accepted.

Open Set Insect Classification using DNA Barcodes (Total Score: 200 points)

In this assignment you are given DNA nucleotides and their corresponding embeddings for $\sim$22,000 insect instances of over 1,000 different species [1]. The data set is obtained from the Barcode of Life Data System (BOLD) [2]. The data is split into two as train and test. Train samples have nucleotides, their matching 1000-dimensional vector embeddings, and species and genus labels whereas testing samples have only nucleotides and their matching vector embeddings. All embeddings are obtained by convolutional neural net (CNN) trained with the train samples. The testing set contains samples from species not represented in the training set. The goal of this assignment is to predict species labels for testing samples represented in the training set and predict genus labels for testing samples not represented in the training set. In other words your classifier will predict the most fine-grained labels available for test samples. If a test sample originates from a species represented in the training set and your classifier accurately predicts the genus label for that sample this will be counted as a misclassification because a more fine-grained, i.e., specific, label was available for that sample.

Descriptions of variables:

gtrain: This is a column vector of size 16128. This variable contains genus level labels for each insect instance in the training set. You can think of these as the parent nodes of the leaf nodes in a tree, where leaf nodes are the species and parent nodes are the genera. All instances with the same gtrain value share the same genus.

ytrain: This is a column vector of size 16128. This variable contains species level labels for each insect instance in the training set. All insect instances with the same ytrain value belong to the same species. For example if you want to find out the row ids of insects that belong to species 1260 you can run $find(ytrain == 1260)$ in Matlab.

nuc_train: This is a struct array of size 16128. This variable contains nucleotide sequences for each insect instance in the training set. Although some sequences might have different lengths most of them have a length of 658. Each sequence is composed of 4 letters A, G, T, C. The symbol $-$ and N (if any) are used to indicate missing values.

[1] download the data file insect_data_csci481_wo_testlabels.mat

[2] http://www.boldsystems.org/

If you want to find out all nucleotides that belong to species 1260 you can run $nuc\_train(ytrain == 1260)$ in Matlab.

emb_train: This is a 2D matrix of size 16128x1000. Each row in this matrix is a high dimensional encoding (or embedding) of the corresponding nucleotide sequence in the training set. In other words you can consider each row of this matrix as a feature vector offering high level characterization of the nucleotide sequence. A deep learning algorithm is used to convert each nucleotide sequence into a numeric vector. This algorithm first converts a nucleotide sequence into a 2D image and then converts the image into a numeric vector of size 1000. These conversions are made by a series of image convolution operators that uses filters of various sizes. These filters and other parameters of the network are optimized to reduce classification error [3].

nuc_test: This is a struct array of size 5989. This variable contains nucleotide sequences for each insect instance in the testing set.

emb_test: This is a 2D matrix of size 5989x1000. Each row in this matrix is a high dimensional encoding (or embedding) of the corresponding nucleotide sequence in the test set. These are obtained using the same network as the one used to obtain emb_train.

Approach: You are allowed to use any classifiers we have covered (or will be covering) in the class such as KNN, Naive Bayes, Bayes classifier, SVM, LDA, and Random Forest. You are highly encouraged to extend these classifiers to handle species and genus classification at the same time.

Submission: You are going to submit a spreadsheet with predicted labels. The order of predictions should match the order of the samples in the $xtest$ matrix. You can include up to three columns in this spreadsheet, one column for each of the different approaches you have tried. You will also send a PDF document explaining each approach in detail. Although you are not required to submit your code your submissions should be reproducible and we may ask your code if needed.

Evaluation: Your submission will be evaluated on two fronts. All submissions will be ranked in terms of the harmonic mean of species, i.e., seen, and genus, i.e., unseen, class accuracies. The accuracy for a class will be computed by $\frac{TP}{TP+FN}$, where TP denotes the number of true positives and FN denotes the number of false negatives. Once accuracies for all classes are computed they will be averaged within each group (species and genus) and the harmonic mean of these two average accuracies will be computed. Top performer will receive the top score of 100 points. The rest of the submissions will be scored based on how well they do relative to the top performer. The remaining 100 points will be assigned based on the novelty/rigor of your approach and the clarity/organization of your write-up.

[3] You don't need to know the technical details of this encoding process to perform well in this competition. If you still need to learn about convolutional neural nets you can check this tutorial https://www.simplilearn.com/tutorials/deep-learning-tutorial/convolutional-neural-network and many other similar tutorials available on the web.