BINF 610 Applied Machine Learning (Spring 2023)

HW2

Due: March 24, 2023

Classification – SVMs, Decision Trees, and Random Forests

In this assignment, you are asked to train classifiers to predict residue contact in protein structures.

The data set contains 10,000 examples, each example is a residue pair, represented by a 375 dimension vector of features built from protein sequence co-evolution, which is believed to provide relevant information regarding whether a residue pair is contact or not.

Task 1: build a SVM for classifying whether a residue pair is contact or not. Use 10-fold cross-validation, train and test a SVM with the following variations:

a) Three different kernels: linear, polynomial of degree 2, and RBF.
b) For each kernel, try three different penalty C= 0, 0.01, and 0.1

Report average ROC score for each case. Plot ROC curve for the case of RBF kernel with C=0.01. Discuss the effects of variations on the performance.

Task 2: build a decision tree (CART algorithm) for classifying whether a residue pair is contact or not. Again, use 10-fold cross-validation to evaluate the performance. Report precision/recall, and discuss the effect of regularization on performance.

a) Full tree without regularization
b) Max-height = 5

Task 3: build a random forest for classifying whether a residue pair is contact or not. Again, use 10-fold cross-validation to evaluate the performance. Report precision/recall, and report the top five features ranked by their importance.

Instruction for submission:

1. All python code should be in a separate file
2. The writeup should be in PDF file.