Homework 2 Report:
Applying kNN Classifier to predict drug sensitivity based on gene expression profiles

Ibrahim Irfanullah, Ahmed Shahkhan

1) kNN Implementation

Data was downloaded and imported into Python (which was used for all analysis in this assignment). Since comma separated value files are easier to parse than text files, the *DREAM_data* text file was converted to a csv file.

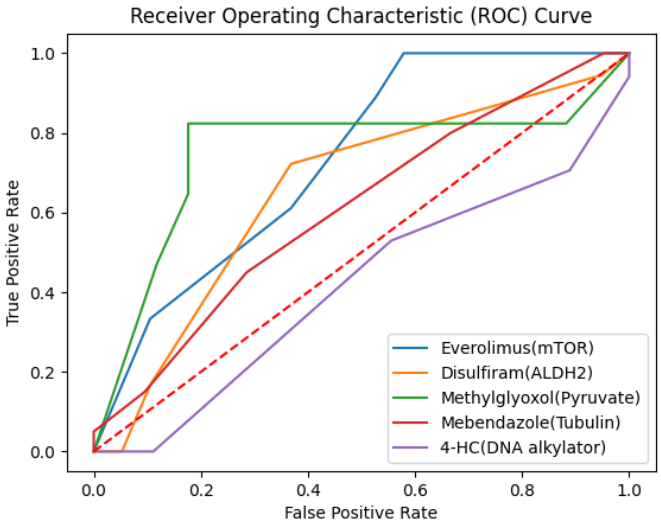
A dynamic kNN classifier on the data was implemented, using the Pearson correlation coefficient as the similarity metric between cell lines' expression profiles. The model is able to take in a cell line expression profile as input and produce a prediction score indicating how likely the patient cell line expression profile is sensitive or resistant to any of the 5 drugs listed in *DREAM_data.*

These prediction scores for each cell line were calculated by taking the fraction of the k-nearest neighbors that are sensitive to the corresponding drug. Consequently, a 46 (samples) by 5 (drugs) matrix was generated with all these prediction scores.

2) kNN Performance Evaluation

Part A: Random Classifier versus kNN Classifier Analysis

Leave-one-out cross-validation was applied to measure the performance of the classifier using a k value of 5. An ROC curve was plotted for each of the 5 drugs. The *False Positive Rates* are on the x-axis and the *True Positive Rates* are on the y-axis. The random classifier is the red-dotted line.



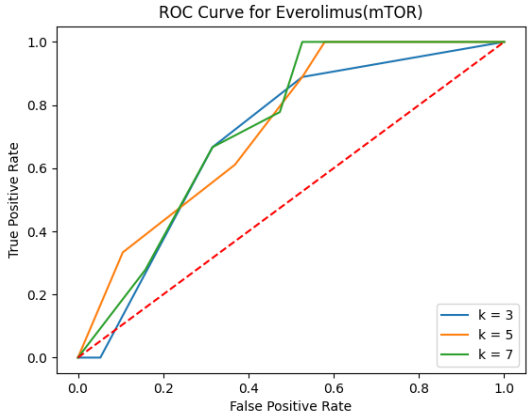
The kNN classifier works better than a random classifier for 4 drugs: Everolimus (mTOR), Disulfiram (ALDH2), Methylglyoxol (Pyruvate), and Mebendazole (Tubulin). As can be seen in the graph below, the ROC curves for these 4 drugs are above the random classifier red-dotted line. In other words, for the majority of values these 4 drugs' have higher true positive rates and lower false false positive rates than a random classifier. The drug 4-HC (DNA alkylator) has less area under its curve than the random classifier, and thus works worse than the random classifier.

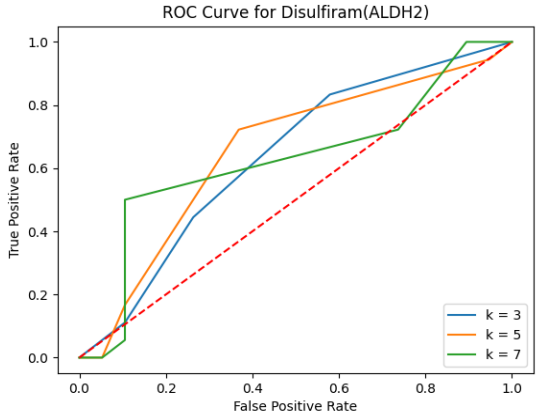
Part B: Best Drug Classification Performance

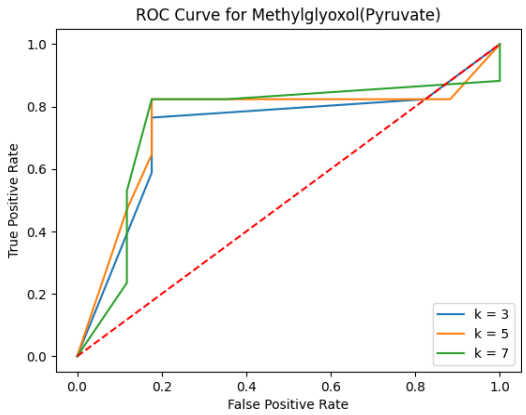
The drug that works best is determined by accurately identifying patient samples who are sensitive to the drug, while minimizing the number of patient samples who are incorrectly classified as sensitive. The drug that does this best is Methylglyoxol because it is correctly identified for patient samples 82 percent of the time and minimizes the number of patient samples that are incorrectly classified to 18 percent of the time. Although Everolimus is correctly identified 100 percent of the time for some values, the ROC curve also indicates the false positive rate is 58 percent. Minimizing the false positive rate is important because it is dangerous to incorrectly classify more than half of the patients as sensitive to a drug (in this case, Everolimus). Therefore, Methylglyoxol was deemed to be modeled best with the kNN approach using a k value of 5.

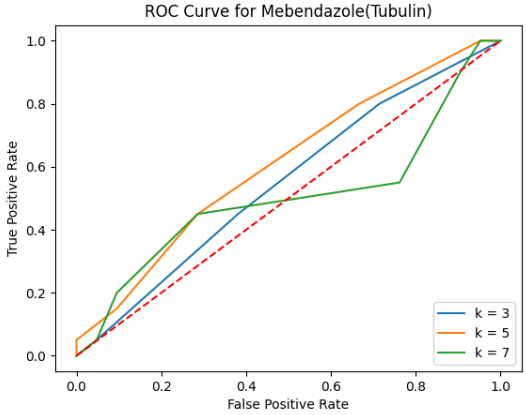
3) Exploration of parameters affecting kNN performance

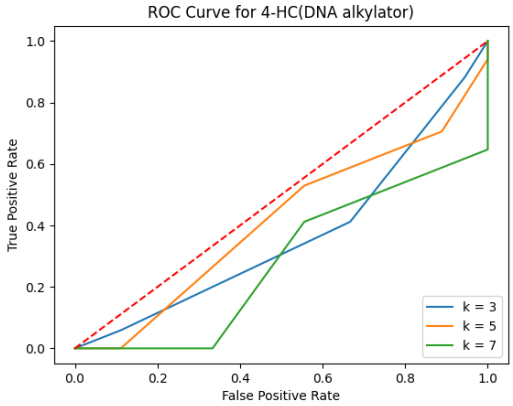
Part A: Classification Results with k=3,5,7











The choice of k does affect the performance of the classifier for certain drugs. For k values 3, 5, and 7, the ROC curve largely remains the same for drugs Everolimus and Methylglyoxol. Additionally, the ROC curves are above and perform much better than the random classifier.

The variability in the ROC curve trajectory increases as the k value increases for drugs Disulfiram, 4-HC, and Mebendazole. A k value of 5 performs slightly better than a k value of 3 as the ROC curve for all 3 of these drugs covers more area. A k value of 7 performs the worst out of the three k values for these drugs.

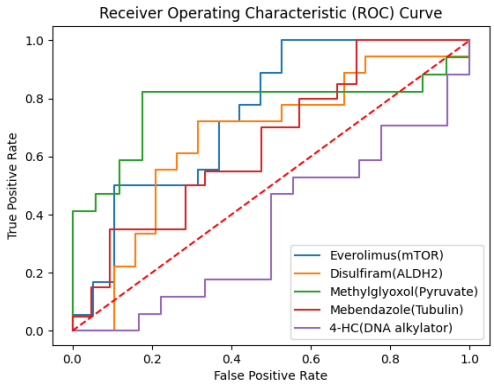
In conclusion, while the variability in the ROC curve trajectory increases as the k value increases for drugs Disulfiram, 4-HC, and Mebendazole, the ROC curve largely remains the same for drugs Everolimus and Methylglyoxol.

Part B: k=5 with new Weighted Score

The following weighted score is used to recalculate the prediction scores:

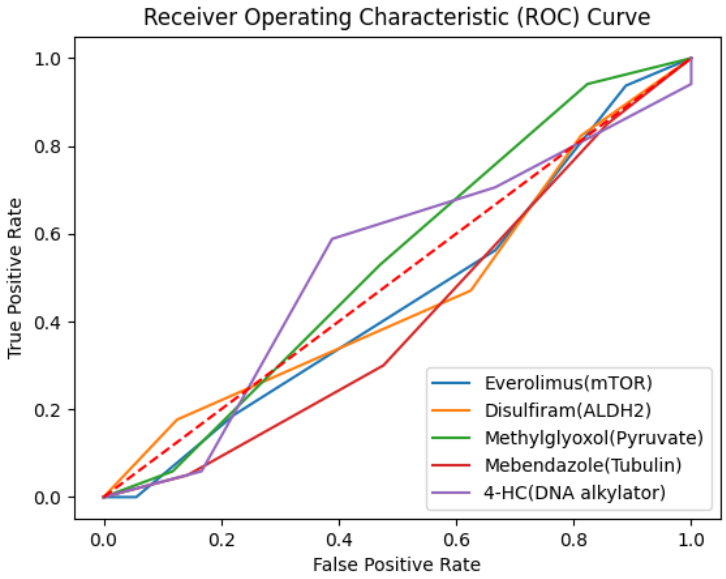
$$S(x) = \sum_{i=1}^{k} sign(y_i) * PCC(x, y_i)$$

In this new equation, y_i represents the nearest neighbors. *sign(y_i)* is +1 for drug-sensitive cell lines or -1 for drug-resistant cell lines. *PCC(x, y_i) is the* Pearson Correlation Coefficient calculated from the kNN implementation. New ROC graphs are plotted below for each drug using these newly calculated weighted scores.

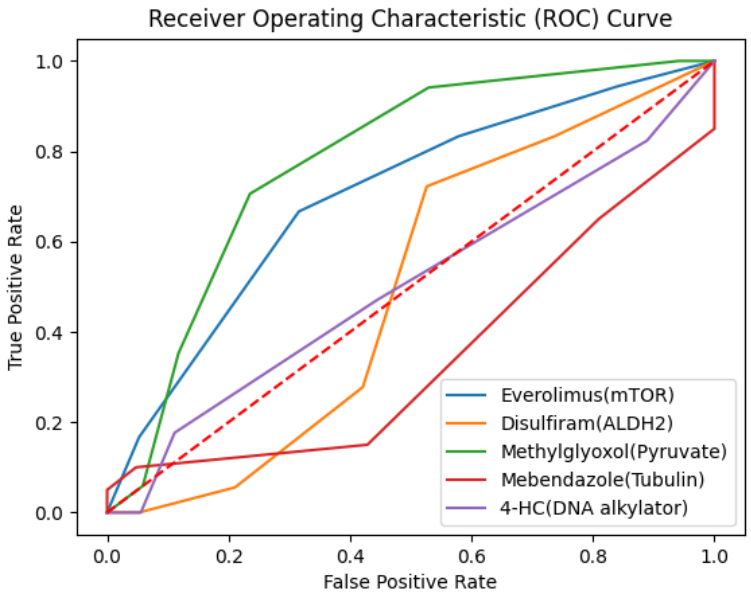


Extra Credit #2

The classification was repeated using the data file RNAseq_quantification.txt, which uses quantified RNAseq data. The results (shown below) indicate that this dataset is less useful than the DREAM dataset in predicting drug sensitivity for each sample.



The two datasets (DREAM and RNAseq_quantification) were then concatenated to determine if simultaneously looking at both datasets could improve the performance of the kNN classifier. The results are shown in the figure below:



As can be seen in the figure, combining the datasets improved the results for some drugs, but made it worse for others.