Project 2

Note: The project report should be up to approximately 3 pages long including relevant figures or tables. Please provide details about methodological aspects, parameters used, rationale, references, and interpretation of the results. In addition to the report, upload in Canvas your python code(s) and their output, and any relevant additional figures. Partial credit will be given to incomplete solutions to each problem.

Identification of the Cellular Localization Sites of Proteins in Ecoli

The study described in the attached paper uses a number of attributes of proteins for predicting protein localization sites in Gram-Negative Bacteria. The number of attributes for the E.coli dataset is 8 (**7** predictive, 1 name). The information about the attributes is:

1. mcg: McGeoch's method for signal sequence recognition.
2. gvh: von Heijne's method for signal sequence recognition.
3. lip: von Heijne's Signal Peptidase II consensus sequence score.
4. chg: Presence of charge on N-terminus of predicted lipoproteins.
5. aac: score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.
6. alm1: score of the ALOM membrane spanning region prediction program.
7. alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence.
8. class

Class in the training dataset consists of **8** different localization sites:

cp (cytoplasm)
im (inner membrane without signal sequence)
pp (perisplasm)
imU (inner membrane, uncleavable signal sequence)
om (outer membrane)
omL (outer membrane lipoprotein)
imL (inner membrane lipoprotein)
imS (inner membrane, cleavable signal sequence)

Class in the test dataset is

un (unknown)

Use the scikit-learn methods to determine a classification methodology of the *training* dataset obtained from these studies, which will result in clusters closely matching the actual localization sites for each entry. Use this methodology to assign the proteins with unknown location from the *test* dataset to the appropriate clusters (i.e., to predict their cellular locations). **The 2 datasets to be used for this project are available in Canvas.**

1. Start with the training dataset.

A. (10 points) Print information about the data contained in the training Ecoli dataset, such as the header, description or summary and graph the dataset using a pair plot representation. The diagonal plots should represent probability distributions. Include the regression line in each plot. The upper triangle should use the scatter plot representation and the lower triangle kernel density estimates.

B. (20 points) Use the **Kmeans method** to cluster the data.

> Start by determining the optimal number of clusters using the elbow method, the average silhouette score, the Calinski-Harabasz score, and the DBI score. Plot the within cluster sum of squares (WCSS) versus the number of clusters (check up to 15 clusters), as well as the other scores versus the number of clusters. How many clusters are optimal based on this analysis? Is this what you were expecting?
>
> Apply the Kmeans method using the number of clusters from E. Save the cluster labels for all the data points. How well does Kmeans perform in labeling the Ecoli proteins from the training set according to their cellular localization?
>
> Plot the dataset in the (‘mcg’,’alm2’) plane using colors based on the cluster labels. Include the cluster centers.

C. (20 points) Use the **Agglomerative Hierarchical clustering method** to cluster the Ecoli data.

> Start by determining the dendogram which allows us to know the clusters that we want our data to be split to. Plot the dendogram for the *complete* linkage. Use the various scores listed in part B to determine the optimal number of clusters. How many clusters are optimal based on this analysis? Is this what you were expecting?
>
> Apply the Agglomerative Hierarchical method with *complete linkage* and Euclidian distance using the 8 clusters expected. Save the cluster labels for the data points. How well does this method perform in labeling the Ecoli proteins from the training set according to their cellular localization?
>
> Plot the dataset in the (‘mcg’,’alm2’) plane using colors based on the cluster labels.
>
> Repeat the above steps for the Agglomerative Hierarchical clustering method with *average linkage.*

D. (5 points) Which of the 2 types of the **Agglomerative Hierarchical clustering methods** performs best on the training Ecoli dataset? How does this compare with the performance of the **Kmeans** method? Which clustering method performs best overall on the training Ecoli dataset?

2. Use the entire Ecoli dataset

E. (15 points) Using the clustering method that performs best in D. predict the cellular localizations of the 15 proteins from the test Ecoli dataset.

F. (30 points) Use **Decision Trees and Random Forests** to predict the cellular localizations of the 15 proteins from the test Ecoli dataset using the trees trained on the proteins from the Ecoli training dataset:

> Use the DecisionTreeClassifier to create the decision tree for the proteins in the Ecoli training set (use tree induction and pruning to produce the final tree) and graphviz to render the tree. What do you learn from the tree about the proteins dataset? (10 points)
>
> Split the proteins Ecoli training dataset into a training set and a test set. Create a Gaussian Classifier using Random Forests. Train the model using the training set. After training, check the accuracy: how often is the classifier correct? How does this accuracy compare with the precision of the clustering methods from part D? (10 points)
>
> Using RandomForest to determine the feature importance. Are all features in the proteins dataset important? (5 points)
>
> Make the predictions for the 15 proteins in the test Ecoli dataset. How well do these predictions compare with the predictions from part E, which are based on clustering? (5 points)