CS 5310 Data Mining

Lab – Chapter 3 Lazy Learning – Classification Using Nearest Neighbors

Peptides Classification – Service Learning

Two datasets, viz., “peptide\_peak30.csv”, and “peptide\_peak40.csv”, are given. They represent peptides data collected in two different lab conditions. We want to use the kNN classification method to train a model that can distinguish peptides that appear in both conditions from those only appear in one of the conditions.

Firstly, we need to process each dataset using the methods in our Chapter 2 lab, to remove the columns we don’t need, split the “1/K0 Range” column into 3 columns, and normalize the columns with numeric data (Refer to Chapter 2 Lab for details).

We will then mark out the peptides that appear in both datasets, create a boolean variable named “Common” with value True indicating that the peptide appears in both datasets and False otherwise. This “Common” variable will be used as the target feature. Note that two peptides are the same if they have the same name, which is in the “Peptide” or 1st column of the dataset.

Then, we combine two datasets into one (using rbind() function), sample 80% of the data to do the training and the remaining to do the testing.

Perform the following activities in R:

1. Load data in “peptide\_peak30.csv”, and “peptide\_peak40.csv”, into 2 separate data frames in R.
2. Process each dataset in the way we did in the Chapter 2 lab. See the notes above.
3. Compare peptides in 2 datasets and create the target feature to indicate whether the peptide appears in both datasets or only one of them. See the notes above.
4. Combine two datasets into one using rbind() function, randomly sample 80% of the data as the training dataset and the remaining 20% as the testing dataset.
5. Train the kNN model using the training dataset. Note that only numeric features should be used in the training process.
6. Use the kNN model to predict the target feature of the testing dataset.
7. Use CrossTable() function to compare the predicted state values to the saved true values and analyze the results.

Submit your source codes (in a .r file) and a Word document containing the cross table your program creates in Blackboard.