02-425/725 HW6

April 8, 2024

1 Subcellular localization

Consider the set of protein localization features `localization_features.txt`. The first column (ID) is the name of the protein, and the second column (CLASS) represent the location of the protein in the cell. The next 56 columns are different features, based on which we want to predict the location. The first 20 features are amino acid contents, and the rest of the features are described in `psortFeatureDescriptions.html.pdf`. There are 12 possible cellular locations, and the dataset contains information from 2391 proteins.

1. Design a Naïve Bayesian Classification method to classify the location of proteins based on their features. Assume each feature is either discrete or Gaussian. Decide on whether to use discrete or Gaussian model for each feature based on the values of the features and the description. Train your method on the first 2000 proteins to learn the parameters, and test it on the next 391 proteins. What is the accuracy of the method?

2. Design a kNN classification method for this problem, and try various values of k. What is the value of k with highest accuracy?

3. Describe what are the bottlenecks of Naïve Bayesian Classification and kNN classification approaches.