BIL566 PATTERN ANALYSIS
MIDTERM EXAM II

ID          : 
Name        : 
Signature   : 
Grade       : 

Note: The deadline for delivery of midterm exam papers is 21st of December (Monday). The solutions will include source codes of the questions, and a word document explaining the findings. The solutions will be submitted electronically via ceng.anadolu.edu.tr or via email (for students who cannot login to the department web site).

Question 1. (60 pts)

Develop a program that performs SFS (Sequential Forward Selection) feature selection on the largest 4 classes (CYT, NUC, MIT, ME3) of Yeast dataset (http://archive.ics.uci.edu/ml/datasets/Yeast). Divide the dataset into training (60%), validation (20%) and test (20%) parts. Use classification accuracy of the classifier (Bayesian, SVM, etc.) obtained on the validation set as the feature evaluation criterion and find out the best feature subset. Then, employ the selected feature subset to measure the accuracy on the test set. Finally, compare the results with the ones obtained with the full feature set.

Note: You can use Matlab, Java or any other programming language for implementation. You will use only one classifier. The choice of the classifier depends on you.

Question 2. (40 pts)

A text collection consisting of 4 classes is provided in the attachment. Each class has 10 documents. Develop a program to compute Gini Index (GI) values of each unique term within the collection. Then, indicate how many unique terms are present in the collection and list top-10 terms (with its GI score) having highest GI scores. No preprocessing (stemming, stop-word removal, etc.) is required.

\[
GI(t) = \sum_{i=1}^{M} P(t \mid C_i)^2 P(C_i \mid t)^2
\]

where \( P(t \mid C_i) \) is the probability of term \( t \) given presence of class \( C_i \). \( P(C_i \mid t) \) is the probability of class \( C_i \) given presence of term \( t \), respectively.

Note: You can use Matlab, Java or any other programming language for implementation.