Assignment #3

CE 764 Hydroinformatics

Fall 2017.

Maximum marks: 10. Due Date: 11 pm, 14 Oct, 2017.

Email to: peterck05@gmail.com, denzilroy@gmail.com with a copy to arpita567@gmail.com

Load 'cancer_dataset' that comes with Matlab by default. Alternately, use the attached .csv files for other programming platforms. The dataset contains 9 input variables each of which defines characteristics of tumour biopsies. There is a total of 699 biopsy samples. (cancerInputs: 9x699 matrix)

The input variables are:

Clump thickness	Marginal Adhesion	Bland chromatin
Uniformity of cell size	Single epithelial cell size	Normal nucleoli
Uniformity of cell shape	Bare nuclei	Mitoses

The response (cancer Targets: (1) Benign or (2) Malignant) consists of a 2x699 matrix where each row indicates a correct category with a one in either element 1 or element 2. For example, row 1 is for benign tumour, hence 1 for benign and 0 for malignant is marked in row 1.

Objective: To build a suitable classifier using A) Classification tree, B) Support vector machines and C) Artificial Neural Networks to identify malignant (harmful) or benign (harmless) tumour based on the 9 input variables.

Instruction common to all subsections: Divide the given data into two subsets of 549 (training + validation if applicable) and 150 observations, chronologically. Retain the 150 observations for testing the accuracy of the classifier built and comparing different classifiers.

A) CART (3 marks)

i) Write a program to grow a tree to full depth using top-down greedy algorithm on the training data to classify the tumour biopsy into benign or malignant. Write the command to access the total number of nodes in the tree.

ii) Prune the tree to the best level using k-fold cross validation using 10 cross-validation samples. Estimate the misclassification percentage using the pruned tree on the testing data.

iii) Use bagging to create an ensemble of 1000 trees. Make prediction on the testing data using the bagged tree. Does ensemble classification by bagging improve prediction efficiency for the given data?

B) SVC/SVM (3 marks)

i) Build a support vector classifier (linear boundary) for the classification problem defined. What is the misclassification percentage for this SVC model?

ii) Plot a 2D scatter of the training observations with Uniformity of cell size as the ordinate and Bare nuclei as the abscissa. Use different colors for the two classes, and mark the SVC decision boundary.

iii) Build a support vector machine (non-linear boundary) with a radial kernel. Does this outperform the SVC? Justify.

C) ANN (4 marks)

i) Write a program to build a feed-forward neural network with 2 hidden layers and 1 output layer to classify malignant and benign tumours. Use 10 nodes in each hidden layer. Use sigmoid transfer function in both the hidden layers and the output layer. Use the 'traingd' (gradient descent) backpropagation algorithm.

ii) Train the neural network on the training observations. Report the learning rate parameter.

iii) Test the performance of the neural network on the testing data using mse and misclassification percentage.

iv) Compare the performance of the best trained neural network, bagged classification tree and SVM for this classification problem.

-----Violation of honour code will result in penalty------

**********End**********