Department of Engineering/Informatics, King's College London
Pattern Recognition, Neural Networks and Deep Learning (7CCSMPNN)
Assignment: Ensemble Methods

This coursework is assessed. A type-written report needs to be submitted online through KEATS by the deadline specified on the module's KEATS webpage. This coursework considers your "own created" dataset to investigate the classification performance using the techniques of Bagging and Boosting. Some simple "weak" classifiers will be designed and combined to achieve an improved classification performance for a two-class classification problem.

Q1. Create a non-linearly separable dataset consisting of at least 20 two-dimensional dataset. Each data is characterised by two points $x_1 \in [-10, 10]$ and $x_2 \in [-10, 10]$ and associated with a class $y \in \{-1, +1\}$. List the data in a table in a format as shown in Table 1 where the first column is for the data points of class "-1" and the second column is for the data points of class "$+1$". (20 Marks)

Class 1: $y = -1$	Class 2: $y = +1$
(x_1, x_2)	(x_1, x_2)
$\vdots$	$\vdots$
(x_1, x_2)	(x_1, x_2)

Table 1: Dataset of two classes.

Q2. Plot the dataset (x axis is x_1 and y axis is x_2) and show that the dataset is non-linearly separable. Represent class "-1" and class "$+1$" using "$\times$" and '$\circ$", respectively. Explain why your dataset is non-linearly separable. *Hint: the Matlab built-in function* `plot` *can be used.* (20 Marks)

Q3. Design Bagging classifiers consisting of 3, 4 and 5 weak classifiers using the steps shown in Appendix 1. A linear classifier should be used as the weak classifier. Explain and show the design of the hyperplanes of weak classifiers. List the parameters of the design hyperplanes.

After designing the weak classifiers, apply the designed weak classifiers and bagging classifier to all the samples in Table 1. Present the classification results in a table as shown in Table 2. The columns "Weak classifier 1" to 'Weak classifier n" list the output class ($\{-1, +1\}$) of the corresponding weak classifiers. The column "Overall classifier" list the output class ($\{-1, +1\}$) of the bagging classifier. The last row lists the classification accuracy in percentage for all classifiers, i.e., $\frac{\text{Number of correct classifications}}{\text{Total number of samples}} \times 100\%$. Explain how to determine the class (for each weak classifier and over all classifier) using one test sample. You will have 3 tables (for 3, 4 and 5 weak classifiers) for this question. Comment on the results (in terms of classification performance when different number of weak classifiers are used). (30 Marks)

Data	Weak classifier 1	$\cdots$	Weak classifier n	Overall classifier
(x_1, x_2), y	$\{-1, +1\}$	$\cdots$	$\{-1, +1\}$	$\{-1, +1\}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	
(x_1, x_2), y	$\{-1, +1\}$	$\cdots$	$\{-1, +1\}$	$\{-1, +1\}$
Accuracy (%)		$\cdots$		

Table 2: Classification results using Bagging technique combining n weak classifiers. The first row "Data" are the samples (both classes 1 and 2) in Table 1.

Q4. Design a Boosting classifier consisting of 3 weak classifiers using the steps shown in Appendix 2. A linear classifier should be used as a weak classifier. Explain and show the design of the hyperplanes of weak classifiers. List the parameters of the design hyperplanes. After designing the weak classifiers, apply the designed weak classifiers and boosting classifier to all the samples in Table 1. Present the classification results in a table as shown in Table 2. Explain how to determine the class (for each weak classifier and boosting classifier) using one test sample. Comment on the results of the overall classifier in terms of classification performance when comparing with the 1st, 2nd and the 3rd weak classifiers, and with the bagging classifier with 3-weak classifiers in Q.3.

(30 Marks)

Appendix 1: Bagging[1]

Q1. Start with dataset $\mathcal{D}$.

Q2. Generate M dataset $\mathcal{D}_1$, $\mathcal{D}_2$, ..., $\mathcal{D}_M$.

- Each distribution is created by drawing $n' < n$ samples from $\mathcal{D}$ with *replacement*.
- Some samples can appear more than once while others do not appear at all.

Q3. Learn weak classifier for each dataset.

- weak classifiers $f_i(\mathbf{x})$ for dataset $\mathcal{D}_i$, $i = 1, 2, \ldots, M$.

Q4. Combine all weak classifiers using a majority voting scheme.

- $f_{\text{final}}(\mathbf{x}) = \text{sgn}\Big(\sum_{i=1}^{M} \frac{1}{M} f_i(\mathbf{x})\Big)$

Appendix 2: Boosting [2]

- Dataset $\mathcal{D}$ with n patterns
- Training procedure:

[1]Details can be found in Section "Bagging" in the Lecture notes
[2]Details can be found in Section "Boosting" in the Lecture notes

Step 1: Randomly select a set of $n_1 \leq n$ patterns (without replacement) from $\mathcal{D}$ to create dataset $\mathcal{D}_1$. Train a weak classifier C_1 using $\mathcal{D}_1$ (C_1 should have at least 50% classification accuracy).

Step 2: Create an "informative" dataset $\mathcal{D}_2$ ($n_2 \leq n$) from $\mathcal{D}$ of which roughly half of the patterns should be correctly classified by C_1 and the rest is wrongly classified. Train a weak classifier C_2 using $\mathcal{D}_2$.

Step 3: Create an "informative" dataset $\mathcal{D}_3$ from $\mathcal{D}$ of which the patterns are not well classified by C_1 and C_2 (C_1 and C_2 disagree). Train a weak classifier C_3 using $\mathcal{D}_3$.

- The final decision of classification is based on the votes of the weak classifiers.
  - e.g., by the first two weak classifiers if they agree, and by the third weak classifier if the first two disagree.

Marking: The learning outcomes of this assignment are that student understands the fundamental principle and concepts of ensemble methods (Bagging and Boosting); is able to design weak classifies; knows the way to form Bagging/Boosting classifier and knows how to determine the classification of test samples with the designed Bagging/Boosting classifiers. The assessment will look into the knowledge and understanding on the topic. When answering the questions, show/explain/describe clearly the steps/design/concepts with reference to the equations/theory/algorithms (stated in the lecture slides). When making comments, provide statements with the support from the results obtained.

Purposes of Assignment: This assignment goes through the detailed steps of handling classification problem using ensemble methods. You have full control of the datasets which is not the case in real scenarios but allows you to achieve the design easier with a small size of dataset. Through this assignment, it helps you to make clear the concept, working principle, theory, classification of samples, design procedure and multiple-class classification techniques using ensemble methods.