Sindhuja Yerramalla
U00839259.

DataMining

Assignment-6

Q1

a) Assumption.

Bagging - Each sample has probability of $(1-1/n)^n$ of being selected as test data. This decreases the variance in prediction.

Boosting :- Each record is assigned with an equal weight of 1/N (N= no of records). Boosting keeps note of mistakes made by learners when predicting from new learner models. This approach decreases the bias in predictions.

b) Construction process :-

Bagging

step1 :- The dataset will be divided into n bootstrap samples by using sampling and replacement method.

step2 :- A classifier will be designed for each bootstrap samples, which are ~~also~~ the training samples

step3 :- Pass the test bootstrap samples to each and every classifier designed for training samples.

Boosting

step 1 :- Assign weights to each record in the dataset and pass each record to classifier.

step 2 :-

If the record is misclassified increase the weight of the record. If the record is classified correctly then decrease the weight of records.

step 3 :- Pass the records with updated weights to next. classifier and repeat step 2 for T iterations

step 4 :- Pass test record to each and every classifier designed in training phase.

c) Final Aggregation of classificatiers

Bagging :

After passing test data to each classifier the final output will be as follows

problem is classification :

The class of the test sample will be the class with maximum occurences from the O/P of each classifier.

Problem is regression

The output of test samples will be mean (or) median of outputs generated by each classifier.

Boosting :-

After passing test data to each classifier the final output will be as follows

problem ~~is~~ classification :-

The class of test sample will be the class with maximum occurences from the output of each classifier.

problem is regression :-

The output of test sample will be mean (or) median of output generated by each classifier.

Q2

Given,

Total no. of cases = 80

No. of cases M_1 classified as positive = 60

No. of cases negative out of 60 predicted = 12.

No. of cases M_1 classified as negative = 80 - 60

= 20

No. of cases positive out of 20 predicted negatives = 5

Confusion matrix

		Predicted class	
		Yes	No
Actual class	Yes	$T_P = 48$	$F_N = 5$
	no	$F_P = 12$	$T_N = 15$ $(20-5)$

a) Precision (P) $= \frac{T_P}{T_P + F_P} = \frac{48}{48+12} = 0.8.$

Recall (r) $= \frac{T_P}{T_P + F_N} = \frac{48}{48+5} = 0.91$

b) True positivity Rate (TPR) $= \frac{T_P}{T_P + F_N} = \frac{48}{48+5}$

$= 0.91.$ //

False positivity rate FPR $= \frac{FP}{FP + TN}$

$= \frac{12}{12 + 15}$

FPR. $= 0.44$ //

∴ M_1 coordinates of M1 on ROC Curve $= (0.44, 0.91)$ /