Ensemble Learning

CSCI 347, Data Mining

Ensemble Learning

Ensemble learning – multiple learning models are strategically constructed to solve a problem

Types:

- Bagging
- Boosting

Pro	Con
Works surprisingly well	Loss of interpretability

Sources of Errors

Sources of errors:

- bias – error from erroneous assumptions in the learning algorithm
- overfitting – model matches training data closely, but not reality (refer to as "variance"

Bagging Versus Boosting

Randomization		Bagging	
Pro	Con	Pro	Con
	Must modify the learning algorithm	No need to modify the learning algorithm	
Applicable to a greater variety of learners – can be applied even to stable learner			Fails with stable learning algorithms

Example

Banks wants a decision rule regarding loan approval. The loan application provides 13 variables of demographic and socio-economic data and the class variable is binary

1 – applicant is credit worthy

0 – applicant is not credit worthy

https://www.analyticsindiamag.com/primer-ensemble-learning-bagging-boosting/

Example - continued

Prediction must take into account:

- Correct decision results in 35% profit at the end of 5 years
- Wrong decision – false positive results in 100% loss
- When predict 0, bank doesn't incur any loss.

		Predicted	
Actual		**Credit Worthy**	**Non-Credit Worthy**
	Credit Worthy	0.35	0.00
	Non-Credit Worthy	-1.00	0.00

Results

Results:

Classification Algorithm	Accuracy	F1-Score
Decision Tree	76.67%	0.74
Bagging	77.67%	0.77
Boosting	78.67%	0.78

Results

		Predicted	
Actual		**Credit Worthy**	**Non-Credit Worthy**
	Credit Worthy	0.35	0.00
	Non-Credit Worthy	-1.00	0.00

Take profits into account, consider the confusion matrix (also called a contingency table) data for each instance.

	Decision Tree		**Bagging**		**Boosting**	
	CW	**NCW**	**CW**	**NCW**	**CW**	**NCW**
CW	196	14	185	25	186	24
NCW	56	34	42	48	40	50

Calculating the profits for each model:
Decision tree – 0.35*196 – 1.00*56 = 12.60
Boosting – 0.35*186 – 1.00*40 = 25.1