DART: Data Addition and Removal Trees

The Second AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-21)

Jonathan Brophy
jbrophy@cs.uoregon.edu

Daniel Lowd
lowd@cs.uoregon.edu

Goal

Efficiently remove training samples from a random forest model.

Motivation

- Privacy (“Right to Be Forgotten”)
- Save time / resources (eco-friendly)
- Instance-attribution explanations
- etc.

Naive Approach

Training Instances

I1
I2
I3
I4
I5

Train

Random Forest

I2 is requested to be deleted.

Updated Training Instances

I1
I3
I4
I5

Train a new model

New Random Forest

DART

- An RF model that enables the efficient removal of data.

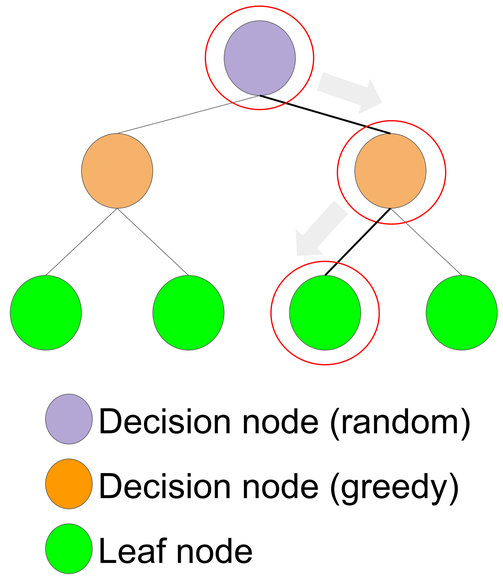
Training a DART Model

- Cache node statistics (*greedy* nodes only).
- Use *random* nodes at the top of each tree.
  - Chooses a feature uniformly at random.
  - DO NOT need to retrain these nodes.
- D-DART: Uses *only* greedy nodes.
- R-DART: Uses a combination of greedy and random nodes.

Deleting a Training Sample

Start at the root node.

1. If random node, move to next affected node and start at 1. Otherwise,
2. Update metadata for each feature split.
3. Recompute split scores.
4. Retrain subtree if a different feature is optimal; if not, move to next affected node and repeat until leaf node, at which point update the leaf value.



Evaluation

- Delete as many samples as possible in the time it takes the naive method to delete 1 sample.
- Samples to be deleted are selected by 2 adversaries:
  - *Random Adversary*: Selects instances uniformly at random.
  - *Worst-of-1000 Adversary*: Picks each sample that causes the greatest amount of retraining out of 1,000 randomly selected samples.

Dataset	n	p
Surgical	14,635	101
Vaccine	26,707	186
Bank Marketing	41,188	112
Adult	48,842	116
Flight Delays	100,000	658
Diabetes	101,766	568
Olympics	206,165	1,016
Credit Card	284,807	150
Census	299,285	414
Synthetic	1,000,000	210
Higgs	11,000,000	100

n: No. samples
p: No. features

Results

- On average, DART is 2-3 orders of magnitude faster than retraining from scratch.
- Higgs dataset
  - Training time: 1.25hr.
  - Avg. deletion time: 0.09s.
  - 50,000 times faster than naive.

Model	Min	Max	Geo. Mean
Random Adversary			
CEDR ($\epsilon = 10$)	1x	21,465x	15x
D-DART	109x	44,228x	677x
R-DART (tol=0.1%)	159x	41,108x	1,500x
R-DART (tol=0.25%)	424x	52,136x	2,713x
R-DART (tol=0.5%)	472x	71,710x	4,037x
R-DART (tol=1.0%)	593x	65,101x	5,979x
Worst-of-1000 Adversary			
CEDR ($\epsilon = 10$)	1x	1,642x	8x
D-DART	10x	2,131x	61x
R-DART (tol=0.1%)	16x	2,446x	184x
R-DART (tol=0.25%)	28x	3,051x	386x
R-DART (tol=0.5%)	32x	6,174x	778x
R-DART (tol=1.0%)	43x	8,185x	1,704x