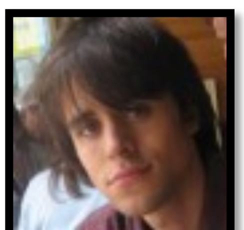
The Boundary Forest Algorithm for Fast Online Learning of High-Dimensional Data

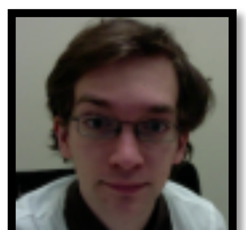
Nate Derbinsky

Assistant Professor, Computer Science and Networking, WIT

Disney Research



Charles Mathy



Jonathan Rosenthal

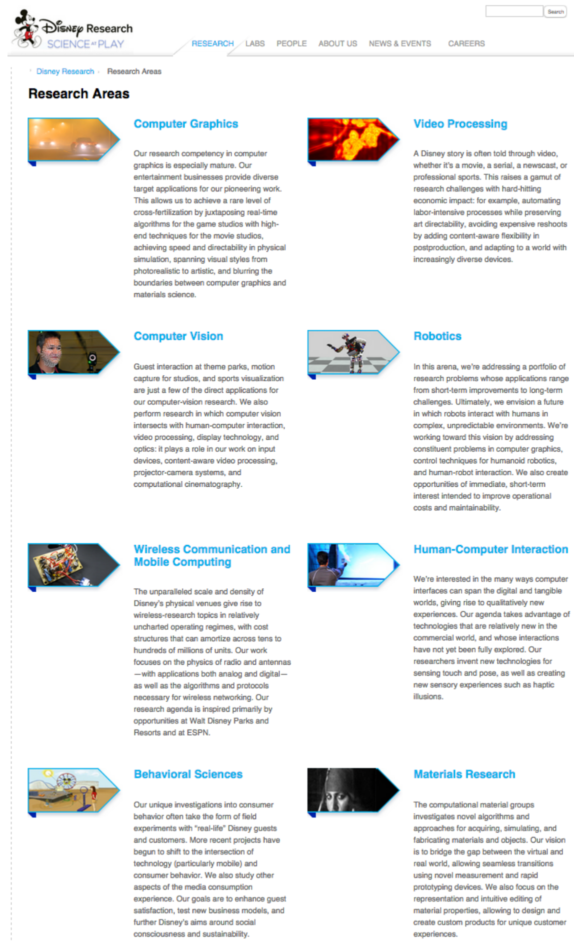
José Bento

Jonathan Yedidia

Outline

1. Disney Research
2. The Problem: Fast Online Learning
3. Boundary Forest Intuition
4. Promising Results: Classification, Regression
5. Algorithm Sketch
6. Evaluation
7. Q&A

Disney Research



Disney Research SCIENCE AT PLAY

Search

RESEARCH LABS PEOPLE ABOUT US NEWS & EVENTS CAREERS

Disney Research › Research Areas

Research Areas

Computer Graphics

Our research competency in computer graphics is especially mature. Our entertainment businesses provide diverse target applications for our pioneering work. This allows us to achieve a rare level of cross-fertilization by juxtaposing real-time algorithms for the game studios with high-end techniques for the movie studios, achieving speed and directability in physical simulation, spanning visual styles from photorealistic to artistic, and blurring the boundaries between computer graphics and materials science.

Video Processing

A Disney story is often told through video, whether it's a movie, a serial, a newscast, or professional sports. This raises a gamut of research challenges with hard-hitting economic impact: for example, automating labor-intensive processes while preserving art directability, avoiding expensive reshoots by adding content-aware flexibility in postproduction, and adapting to a world with increasingly diverse devices.

Computer Vision

Guest interaction at theme parks, motion capture for studios, and sports visualization are just a few of the direct applications for our computer-vision research. We also perform research in which computer vision intersects with human-computer interaction, video processing, display technology, and optics: it plays a role in our work on input devices, content-aware video processing, projector-camera systems, and computational cinematography.

Robotics

In this arena, we're addressing a portfolio of research problems whose applications range from short-term improvements to long-term challenges. Ultimately, we envision a future in which robots interact with humans in complex, unpredictable environments. We're working toward this vision by addressing constituent problems in computer graphics, control techniques for humanoid robotics, and human-robot interaction. We also create opportunities of immediate, short-term interest intended to improve operational costs and maintainability.

Wireless Communication and Mobile Computing

The unparalleled scale and density of Disney's physical venues give rise to wireless-research topics in relatively uncharted operating regimes, with cost structures that can amortize across tens to hundreds of millions of units. Our work focuses on the physics of radio and antennas —with applications both analog and digital— as well as the algorithms and protocols necessary for wireless networking. Our research agenda is inspired primarily by opportunities at Walt Disney Parks and Resorts and at ESPN.

Human-Computer Interaction

We're interested in the many ways computer interfaces can span the digital and tangible worlds, giving rise to qualitatively new experiences. Our agenda takes advantage of technologies that are relatively new in the commercial world, and whose interactions have not yet been fully explored. Our researchers invent new technologies for sensing touch and pose, as well as creating new sensory experiences such as haptic illusions.

Behavioral Sciences

Our unique investigations into consumer behavior often take the form of field experiments with "real-life" Disney guests and customers. More recent projects have begun to shift to the intersection of technology (particularly mobile) and consumer behavior. We also study other aspects of the media consumption experience. Our goals are to enhance guest satisfaction, test new business models, and further Disney's aims around social consciousness and sustainability.

Materials Research

The computational material groups investigates novel algorithms and approaches for acquiring, simulating, and fabricating materials and objects. Our vision is to bridge the gap between the virtual and real world, allowing seamless transitions using novel measurement and rapid prototyping devices. We also focus on the representation and intuitive editing of material properties, allowing to design and create custom products for unique customer experiences.

A Common Problem

Approximate complicated functions
Approximate NN -> Classification, Regression

<u>Requirements</u>

- Incremental
- Fast to train & query
- Scale well given a large number of examples/dimensions

<u>Potential Application Areas</u>

- Real-time learning (e.g. robotics ala RL, vision)
  – Perception, action modeling
- Scalable optimization/simulation

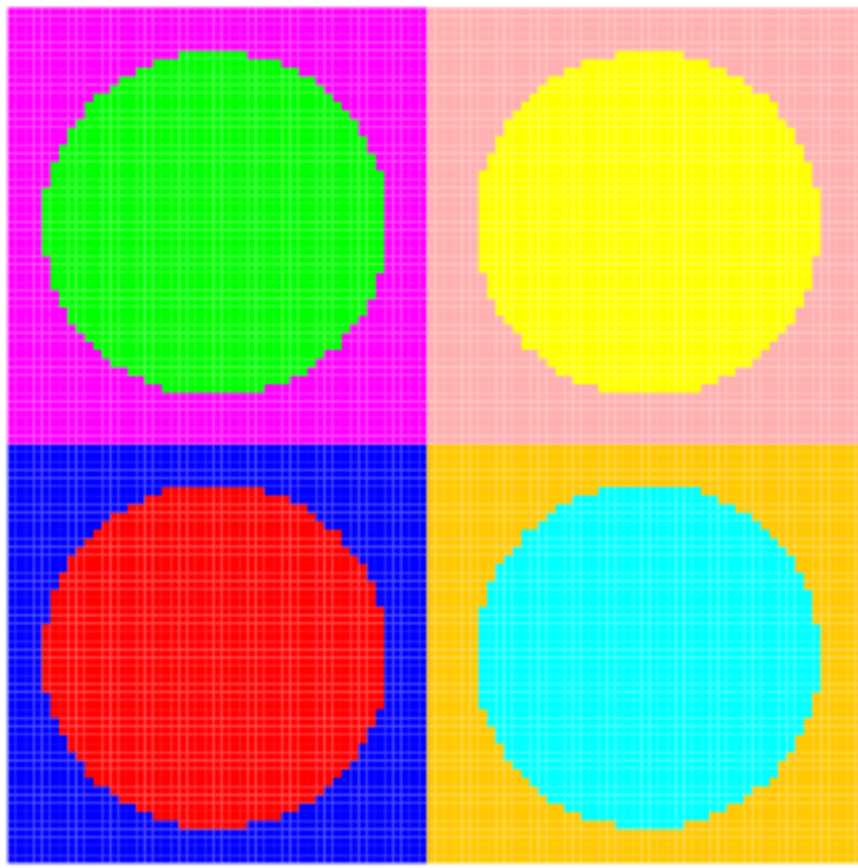
Boundary Forest

Online algorithm that performs effectively and efficiently
- Accuracy: ~kNN
- Time: *O*(logN), both train & query
- Memory: *O*(N)

Ensemble of Boundary Trees, each…

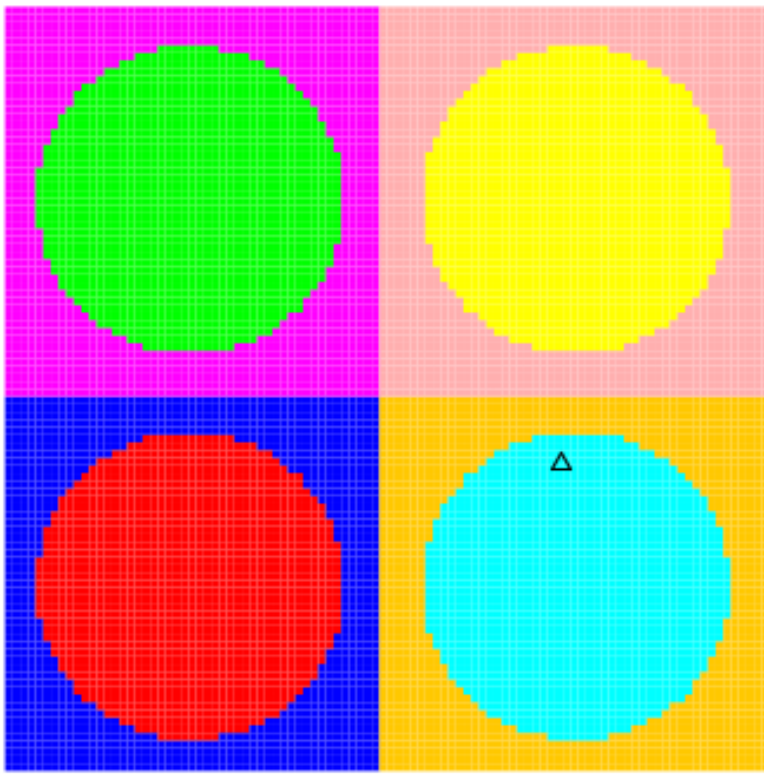
- stores a <u>subset</u> of examples (i.e. instance-based/non-parametric)
  - only those that inform "boundaries" (similar to incremental Condensed NN)
- incrementally builds a <u>graphical search structure</u>
  - queries/trains by **greedily** following/appending-to a search tree w.r.t. distance metric $d(x, y)$

A 2D Classification Example

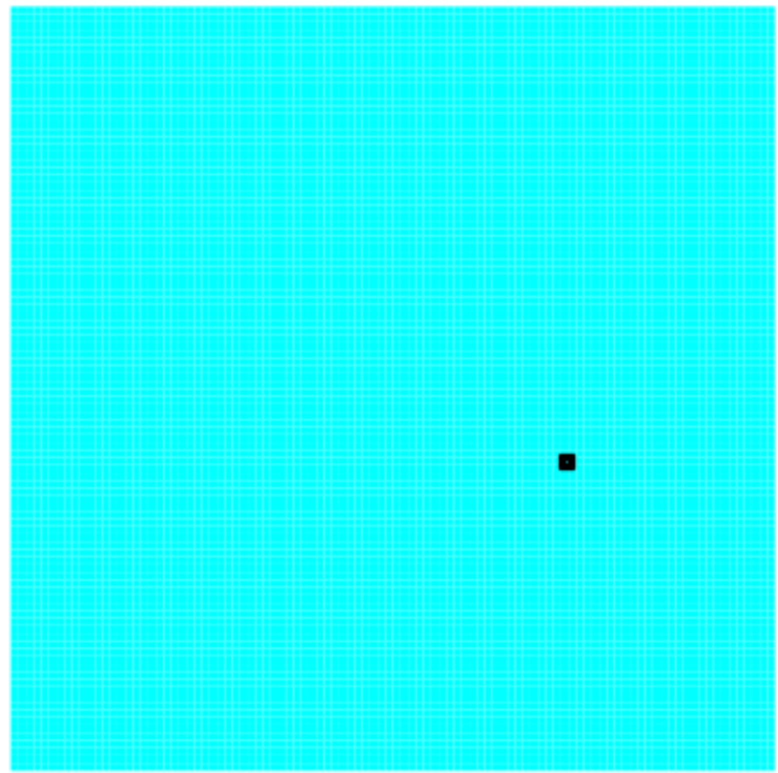


Interleaved Train/Query (1)

Ground Truth

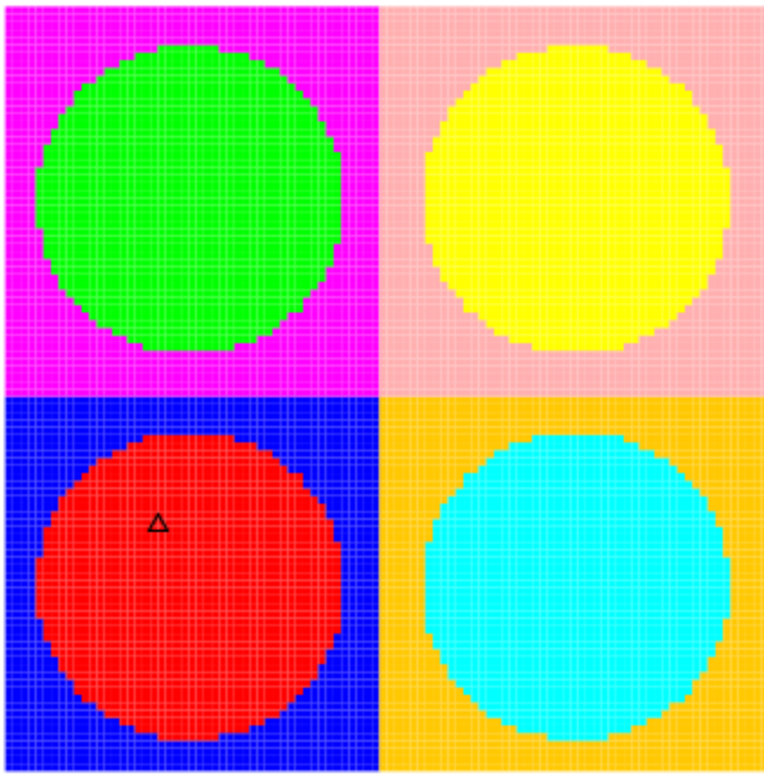


Boundary Tree

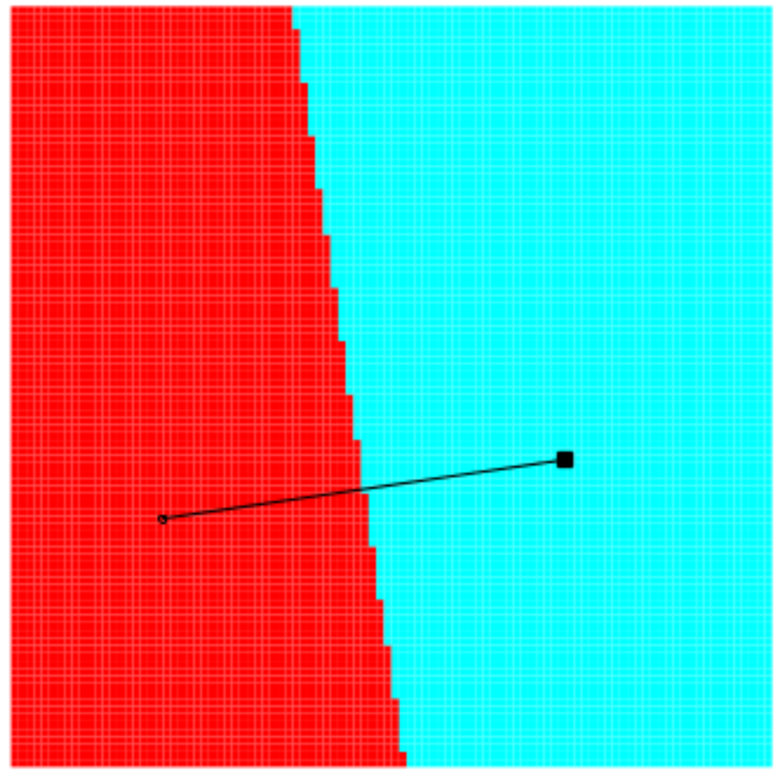


Interleaved Train/Query (2)

Ground Truth

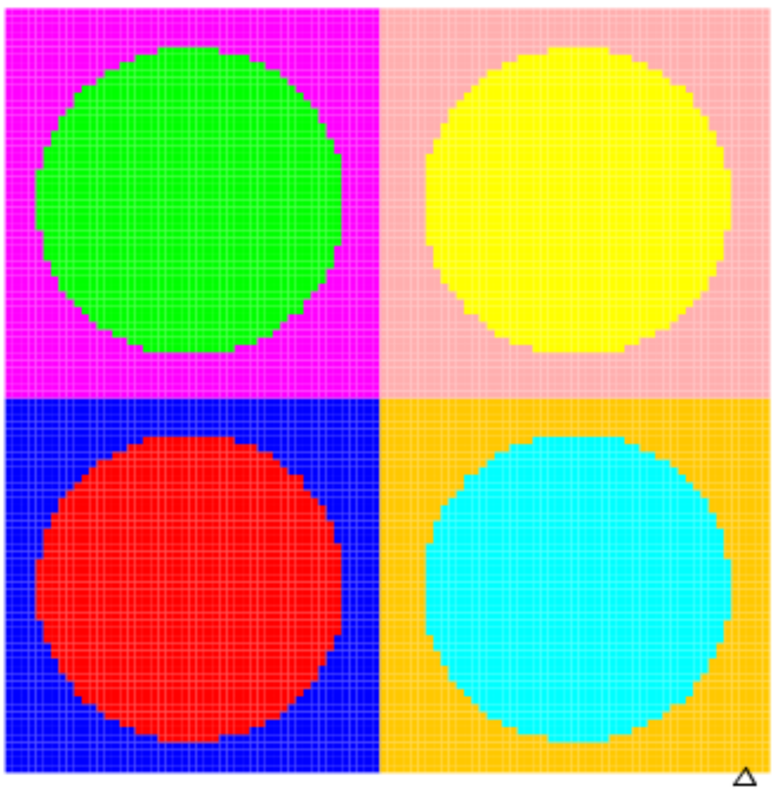


Boundary Tree

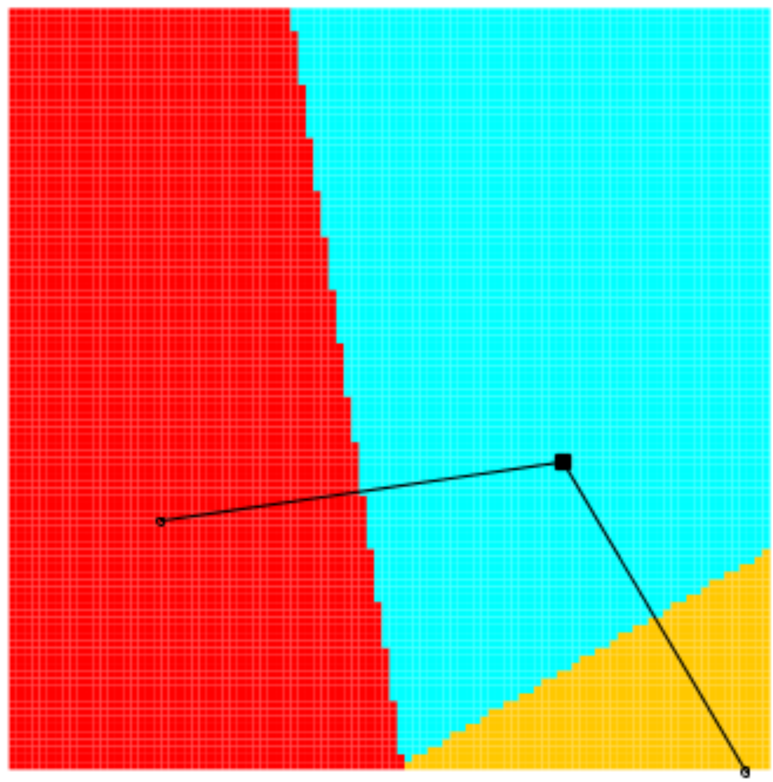


Interleaved Train/Query (3)

Ground Truth

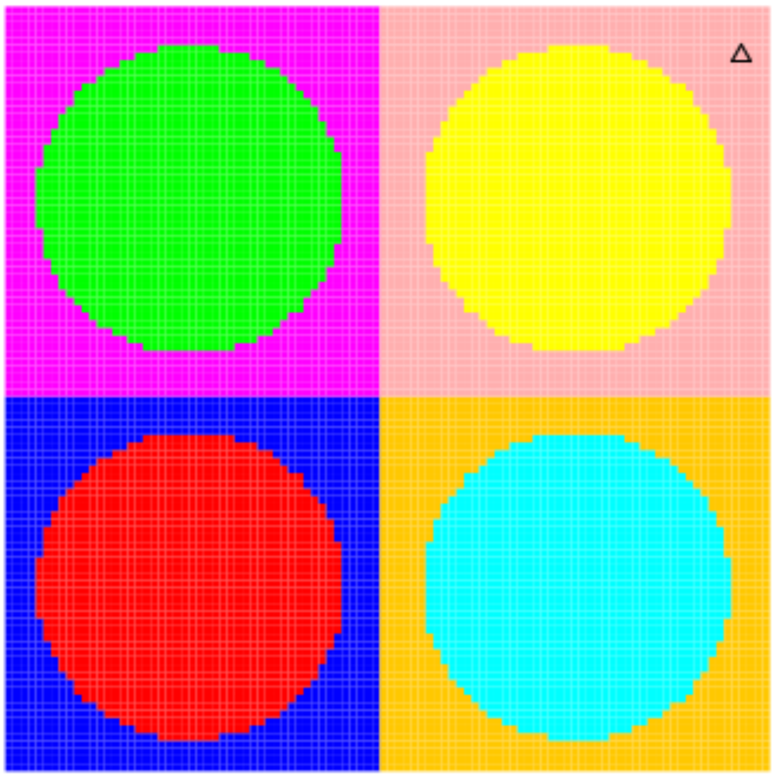


Boundary Tree

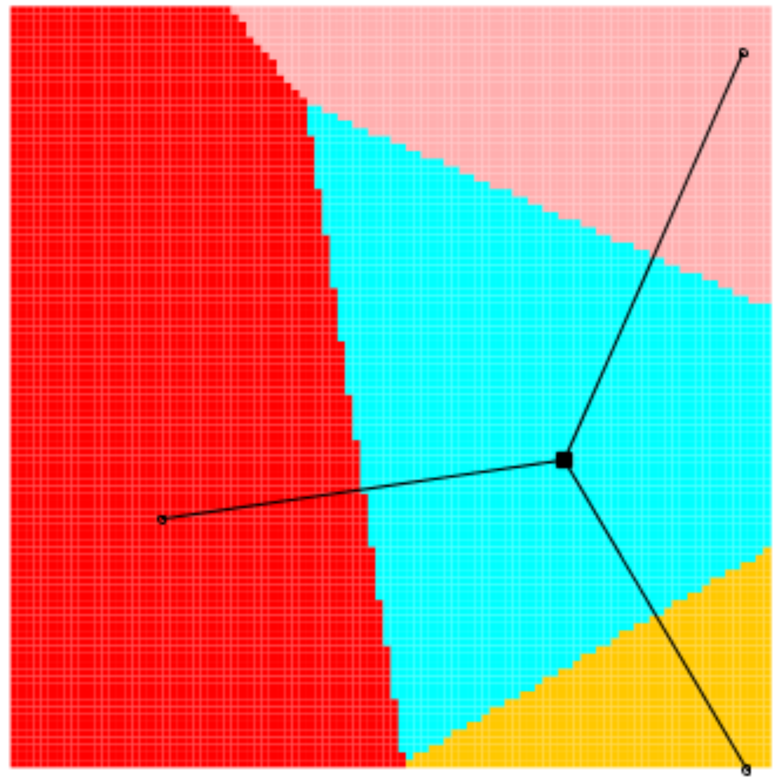


Interleaved Train/Query (4)

Ground Truth

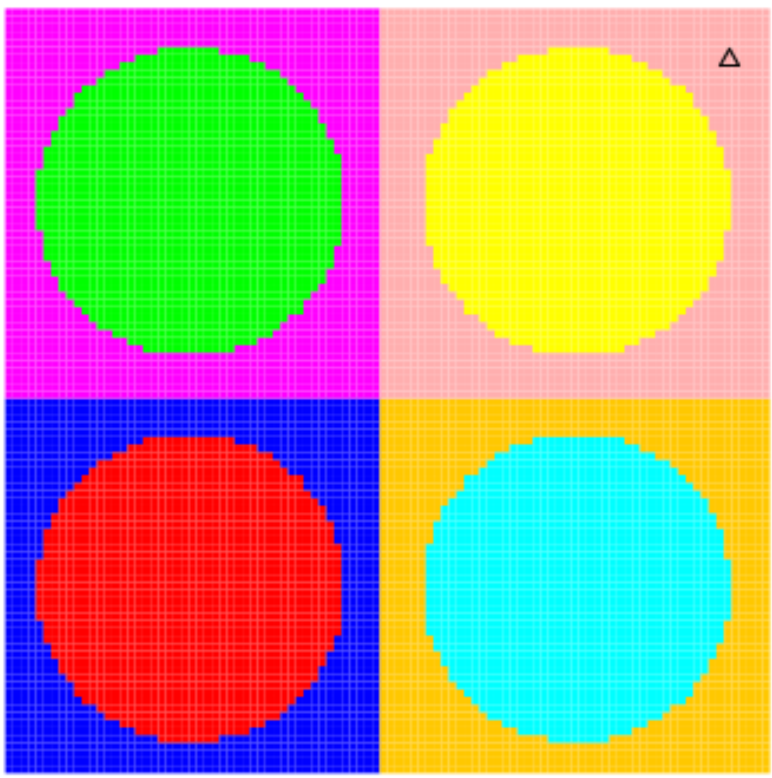


Boundary Tree

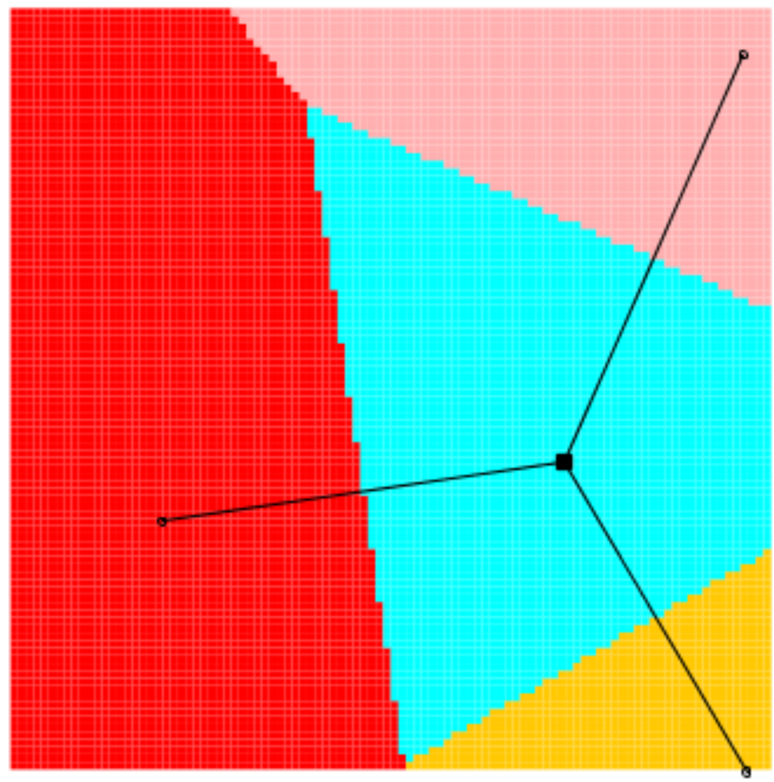


Interleaved Train/Query (5)

Ground Truth

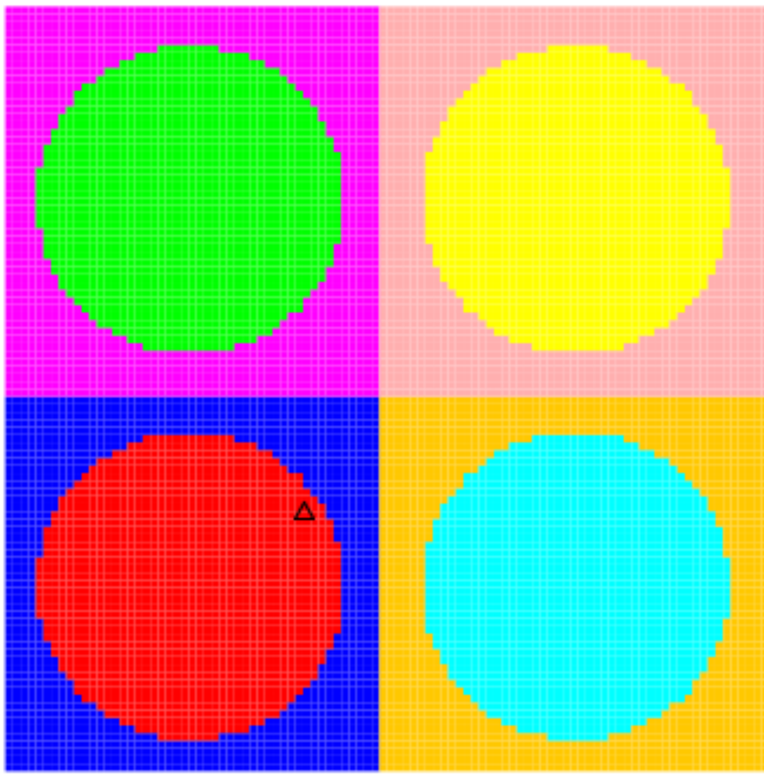


Boundary Tree

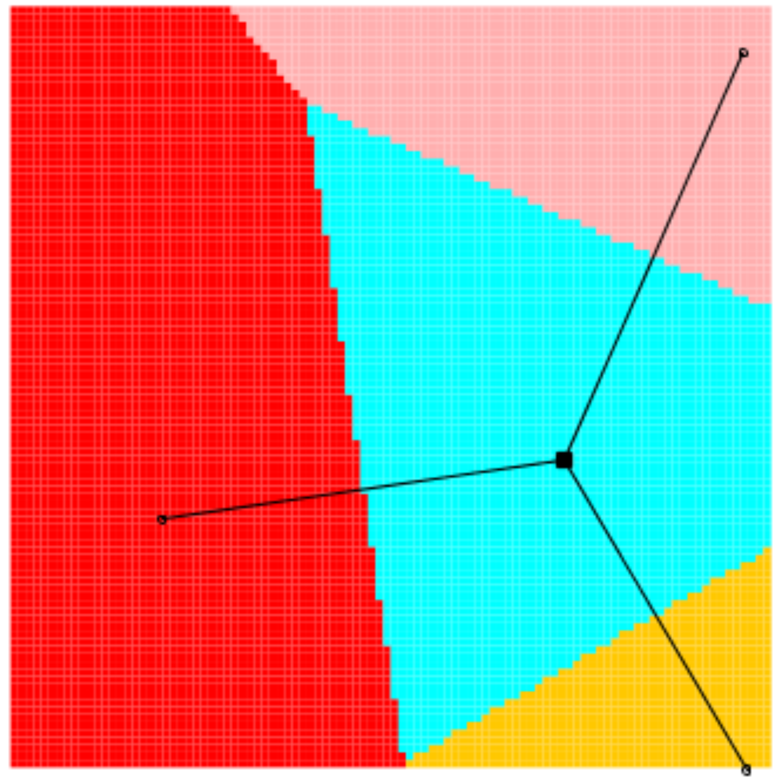


Interleaved Train/Query (6)

Ground Truth

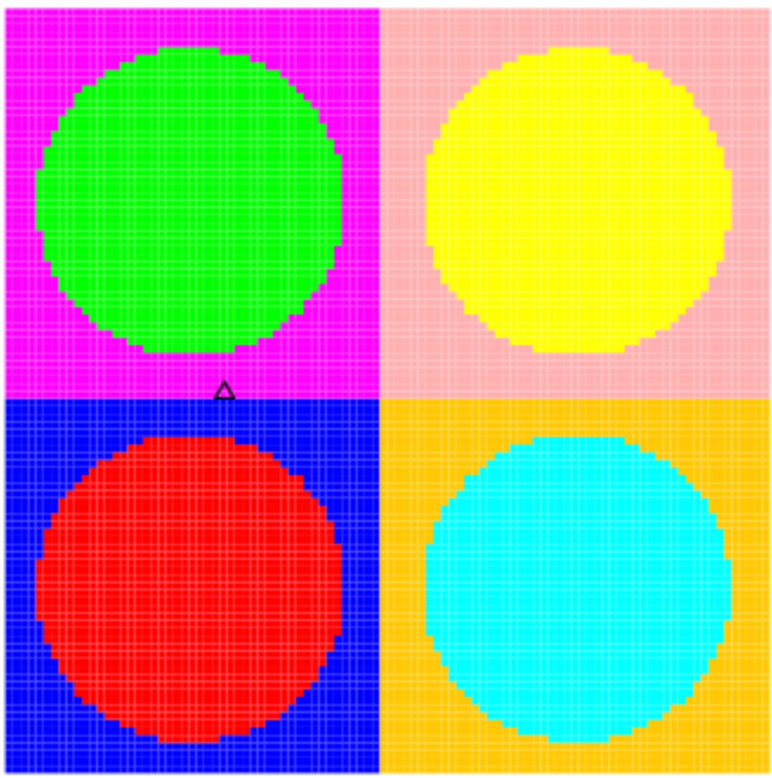


Boundary Tree

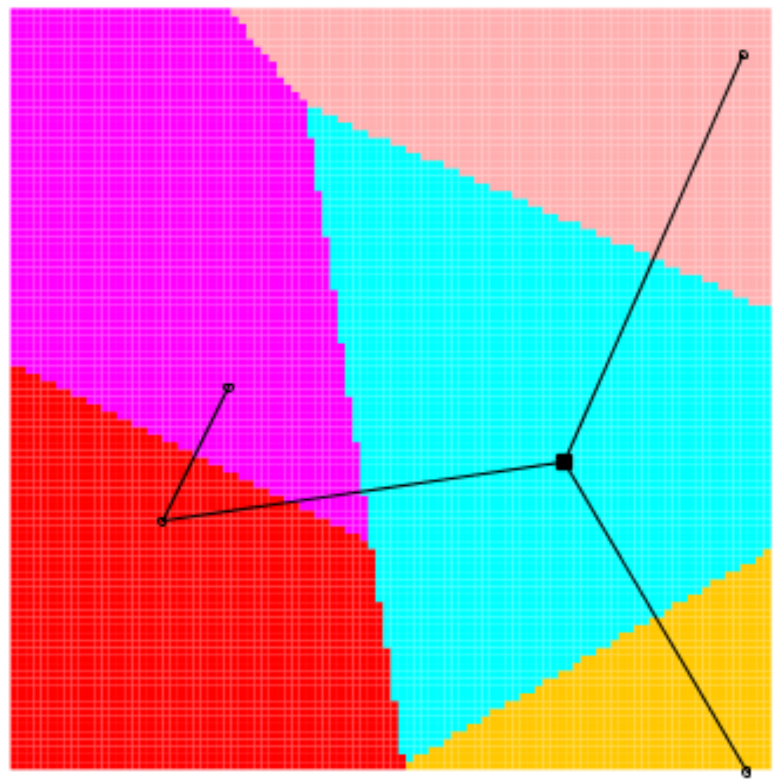


Interleaved Train/Query (7)

Ground Truth

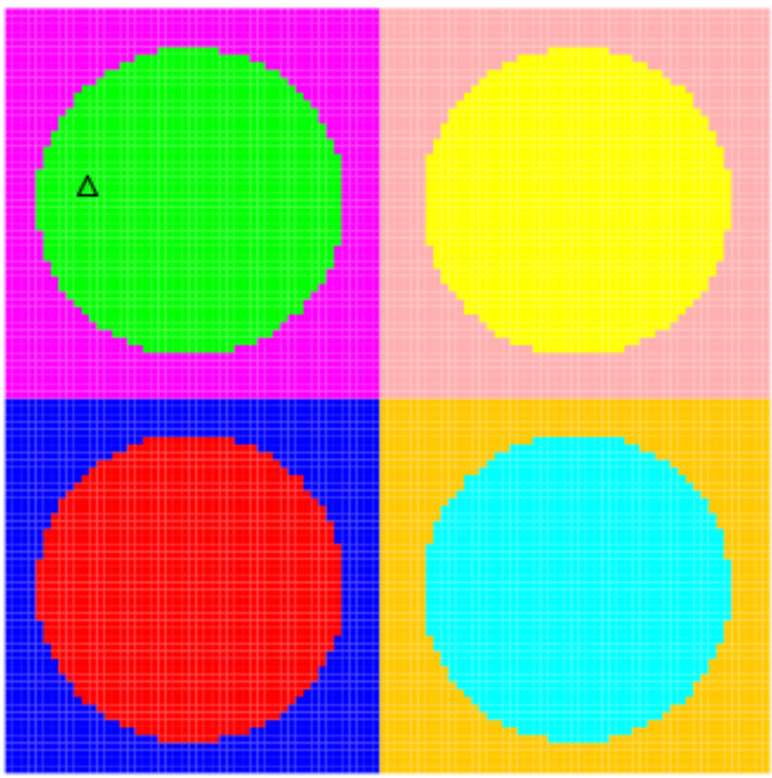


Boundary Tree

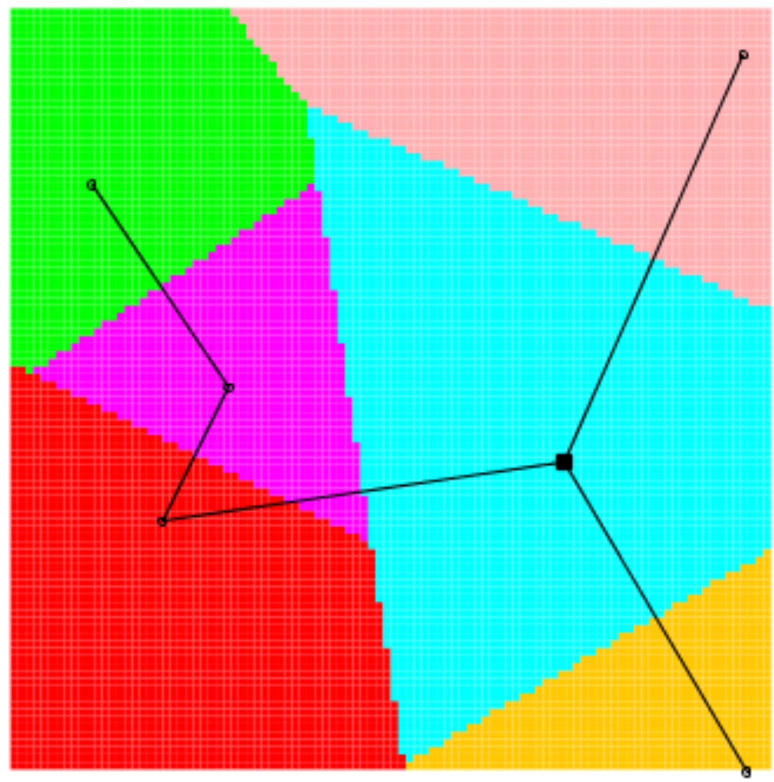


Interleaved Train/Query (8)

Ground Truth

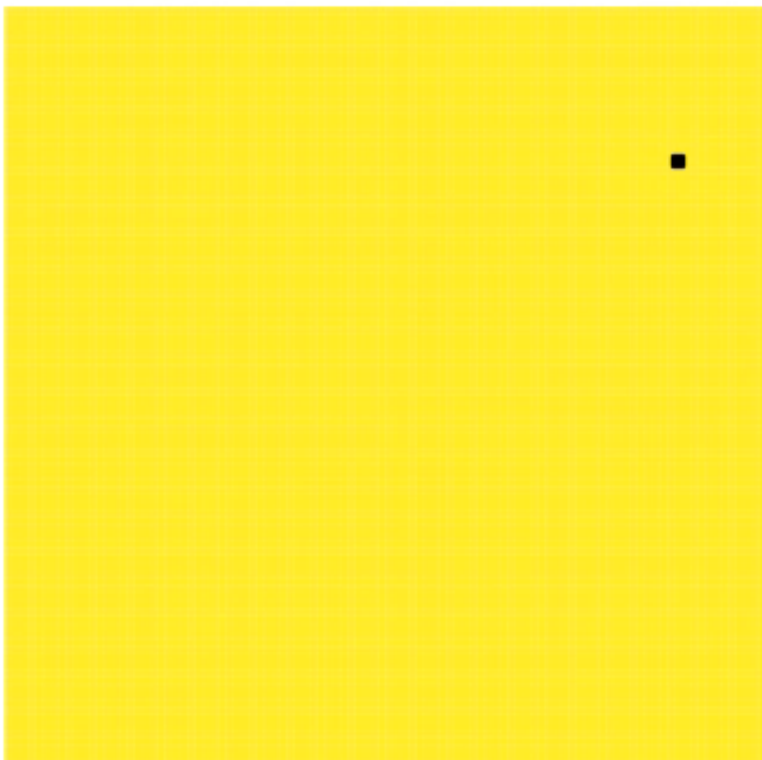


Boundary Tree

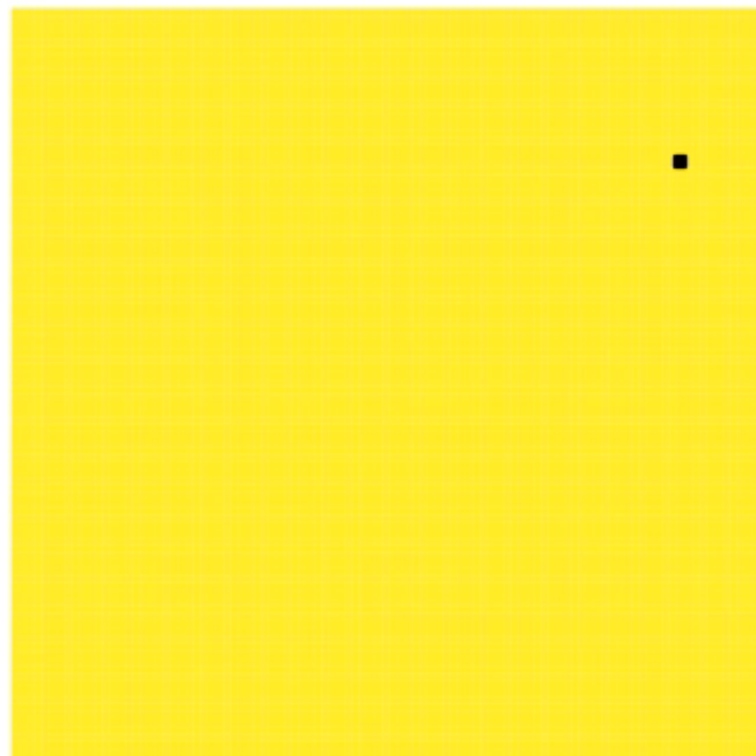


Performance & Scaling

Boundary Tree



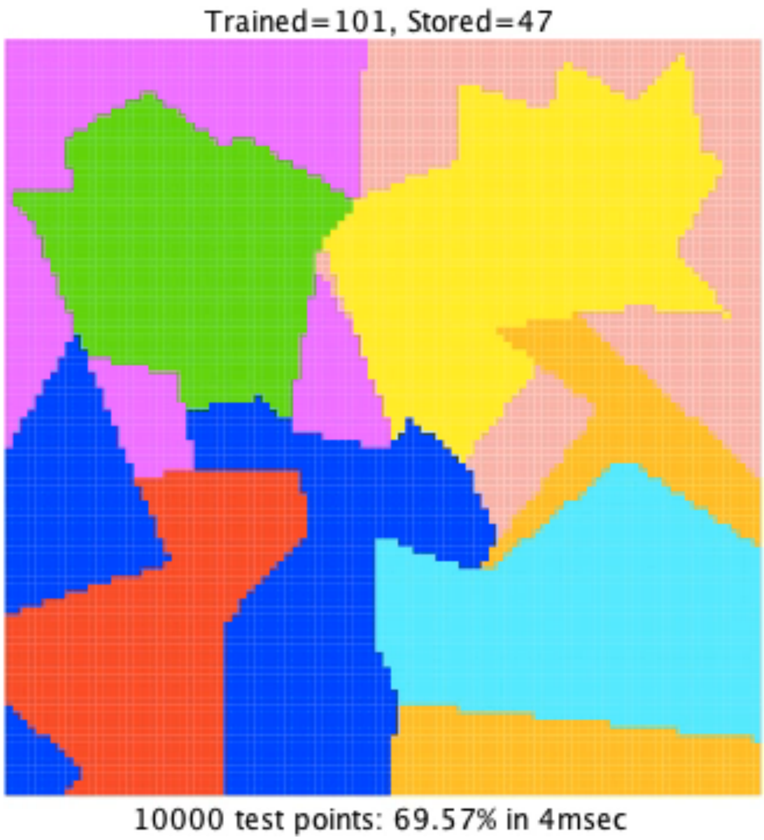
1-NN via Linear Scan



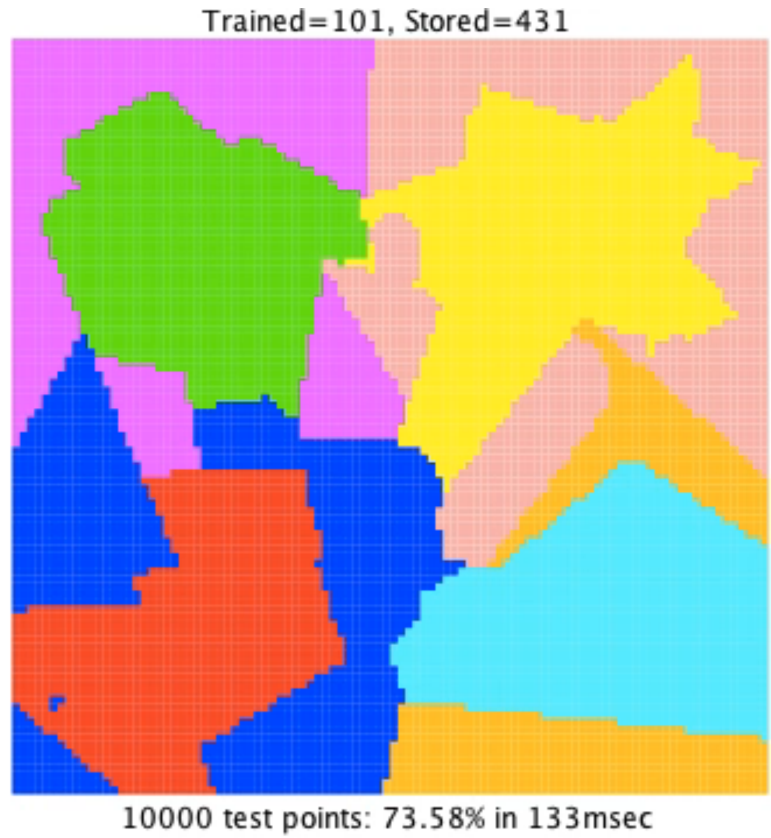
Improving Accuracy via Forests

Linear increase in memory + time

1 Tree



10 Trees



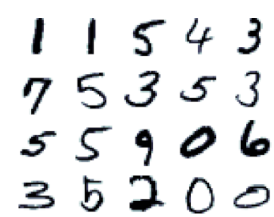
Classification Results

MNIST (60k training, 10k testing, 784 pixels)

Wall Clock Time (seconds)

	Training	Testing	Total
BF(50, 50)	103	2.3	**105.3**
1-NN	0	2900	2900.0
3-NN	0	3200	3200.0
RF(50, 50)	310	0.3	310.3

<2 msec train
<1 msec query



Error, Euclidean Distance

BF(1, 50)	1-CNN	RF(50, 50)	1-NN	3-NN	BF(50, 50)
12.15%	6.70%	3.16%	3.09%	2.83%	**2.32%**

Regression Results
YearPredictionMSD

- 463,715 (training) / 51,630 (testing)
- 90 features
- ~30x faster than 1-NN

RMSE, Euclidean Distance

1-NN	3-NN	BF(50, 50)
14.05	11.59	**10.41**

Algorithm Sketch

Required Parameters

- n_t = number of trees
- k = maximum outdegree
  - Typically leads to eventual logarithmic scaling
- $d(x, y)$ = distance metric
  - Need not be true metric, no assumptions made about properties

Algorithm Sketch
Boundary Tree

Query(y)

- v = root
- loop
  - $cand$ = children(v)
  - if |children(v)| < k
    - $cand = cand \cup v$
  - $v_{min} = argmin_{w < cand}\, d(w, y)$
  - if $v_{min} = v$: break;
  - $v = v_{min}$

Result

- NN: v_{min}
- Classification: class(v_{min})
- Regression: value(v_{min})

Train(y)

- n = Query(y)
- if ShouldAdd(n, y)
  - Connect(n, y)

ShouldAdd

- NN: True
- Classification: Diff. Class
- Regression: Diff. by ε

Algorithm Sketch

Boundary Forest

Query(y)

- for t_i : trees
  - result[i] = t_i.Test(y)

Train(y)

- for t_i : trees
  - t_i.Train(*y*)

Result

- NN: smallest *d*
- Classification: 1/*d* vote
- Regression: 1/*d* average

Initialization

- Root(t_i) = example[*i*]
- *r* = remaining (n_t-1)
  - t_i.Train(Rand(*r, i)*)

Evaluation

- Fast & online algorithm that's easy to code/ understand
- Good performance on classification, regression, a-NN retrieval
- Many potential applications

- Needs a metric; little exploration of dynamic distance functions
- No work yet studying structured/temporal representations
- Future: incorporating dynamic priors

Thank You :)

Questions?

Nate Derbinsky
Assistant Professor
Computer Science and Networking
derbinskyn@wit.edu - DOBBS 140