Hyperbolic Embeddings of Supervised Models


Richard Nock


Ehsan Amid


Frank Nielsen

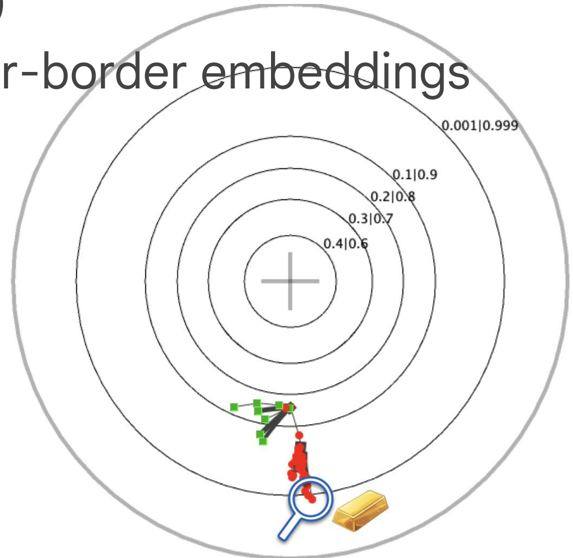

Alexander Soen


Manfred K. Warmuth

Google Research

Summary

- Work about embedding *supervised classifiers* in Poincaré disk model
- Three separate contributions:
  - Embed a single prediction in Poincaré disk model
  - Embed a decision tree DT (or a boosted ensemble)
  - Correct a downside of Poincaré disk model for near-border embeddings
- Code, etc: https://richardnock.github.io/



Embed a single prediction

- In supervised learning, mapping posterior prediction ⟶ real-valued prediction done by (canonical) link ψ of a loss; $|\psi|$= ***<u>confidence</u>***. Ex: log-loss

$$|\psi_{\mathrm{LOG}}(p)| = \log\left(\frac{1+r}{1-r}\right) \quad \text{with} \quad r \doteq |2p-1| \quad \text{and} \quad p = \text{posterior estimation}$$

- 2D Poincaré disk $\mathbb{B}$ = model of hyperbolic geometry (convenient for tree-based representations), with distance to origin of $\boldsymbol{z} \in \mathbb{B}$ given by

$$d_{\mathbb{B}}(\boldsymbol{z}, \boldsymbol{0}) = \log\left(\frac{1+r}{1-r}\right) \quad \text{with} \quad \|\boldsymbol{z}\| = r$$

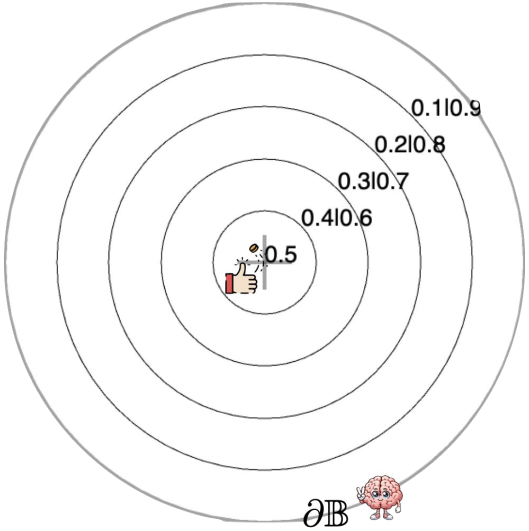
- Suggests embedding a single prediction p by $\boldsymbol{z} \in \mathbb{B}$ with $\|\boldsymbol{z}\| = |2p-1|$

Google Research

– Nock et al., "Hyperbolic Embeddings of Supervised Models", NeurIPS'24

From single to many tree-based predictions (DT)

- (Supervised) benefits of Poincaré disk include: equidistant isolines with respect to prior *p*; center of disk is the worst possible prediction information-wise; the closer to the disk border, the higher the confidence



- A decision tree DT has priors at *each* node so it is natural to want to embed the *full* tree (nodes + architecture)...

Google Research

– *Nock et al., "Hyperbolic Embeddings of Supervised Models", NeurIPS'24*

A direct embedding of a DT is messy

- Indeed, confidences are *not* monotonic from the root to a leaf in general...

```
[#1] -1.6904
├─[#2] -2.5676
│ ├─[#4] -3.0330
│ │ ├─[#7] -2.6730
│ │ │ ├─[#8] -2.9557
│ │ │ │ ├─[#10] -3.5244
│ │ │ │ │ ├─[#12] -5.1468
│ │ │ │ │ │ ├─[#20] -5.8149
│ │ │ │ │ │ │ ├─[#22] -7.0414
│ │ │ │ │ │ │ └─[#23] -4.1897
│ │ │ │ │ │ │   ├─[#140] -4.5850
│ │ │ │ │ │ │   │ ├─[#146] -5.2257
│ │ │ │ │ │ │   │ └─[#147] -2.3026
│ │ │ │ │ │ │   └─[#141] -0.6931
│ │ │ │ │ │ └─[#21] -2.1401
│ │ │ │ │ └─[#13] -3.1053
│ │ │ │ │   ├─[#14] -3.8425
│ │ │ │ │   │ ├─[#24] -3.0445
│ │ │ │ │   │ │ ├─[#62] -5.0370
│ │ │ │ │   │ │ └─[#63] -2.6856
```

$\cancel{|\psi| \nearrow}$

– *Nock et al., "Hyperbolic Embeddings of Supervised Models", NeurIPS'24*

Nice direct embedding of a **Monotonic** DT

- … so we extract its monotonic subtree = *Monotonic Decision Tree* (MDT), where monotonicity is ensured. Then a modified Sarkar algorithm embeds full MDT

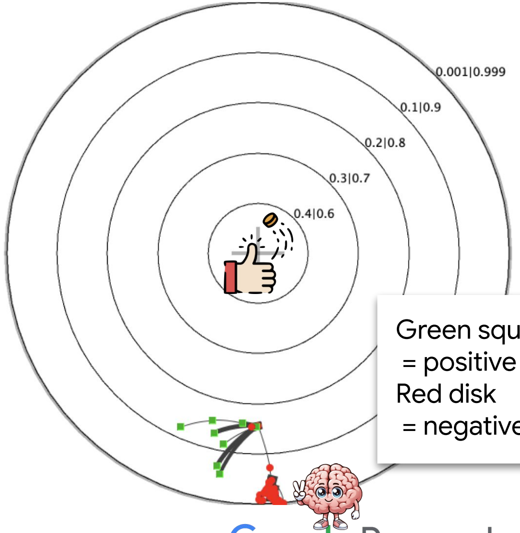
```
[#1] -1.6904
├─[#2] -2.5676
│ ├─[#4] -3.0330
│ │ ├─[#7] -2.6730
│ │ │ ├─[#8] -2.9557
│ │ │ │ ├─[#10] -3.5244
│ │ │ │ │ ├─[#12] -5.1468
│ │ │ │ │ │ ├─[#20] -5.8149
│ │ │ │ │ │ │ ├─[#22] -7.0414
│ │ │ │ │ │ │ └─[#23] -4.1897
│ │ │ │ │ │ │   ├─[#140] -4.5850
│ │ │ │ │ │ │   │ ├─[#146] -5.2257
│ │ │ │ │ │ │   │ └─[#147] -2.3026
│ │ │ │ │ │ │   └─[#141] -0.6931
│ │ │ │ │ │ └─[#21] -2.1401
│ │ │ │ │ └─[#13] -3.1053
│ │ │ │ │   ├─[#14] -3.8425
│ │ │ │ │   │ ├─[#24] -3.0445
│ │ │ │ │   │ │ ├─[#62] -5.0370
│ │ │ │ │   │ │ └─[#63] -2.6856
```

⇨

```
[#1] -1.6904
├─[#2] -2.5676)
│ └─[#4] -3.0330
│   ├─[#10] -3.5244
│   │ ├─[#12] -5.1468
│   │ │ └─[#20] -5.8149
│   │ │   └─[#22] -7.0414
│   │ ├─[#14] -3.8425
│   │ │ ├─[#62] -5.0370}
│   │ ├─[#120] -4.8283
│   │ └─[#25] -5.4582
│   │   └─[#43] -6.4708
│   │ ├─[#48] -3.7377
│   │ └─[#92] -4.1431
│   │   └─[#151] -4.4601
│   │     └─[#168] -5.0106
│   │ ├─[#152] -3.8067
│   │ └─[#172] -4.0342
│   ├─[#82] -3.9120
```

$|\psi| \nearrow$

⇨



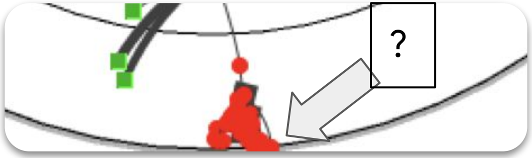
Green square
= positive class
Red disk
= negative class

Google Research

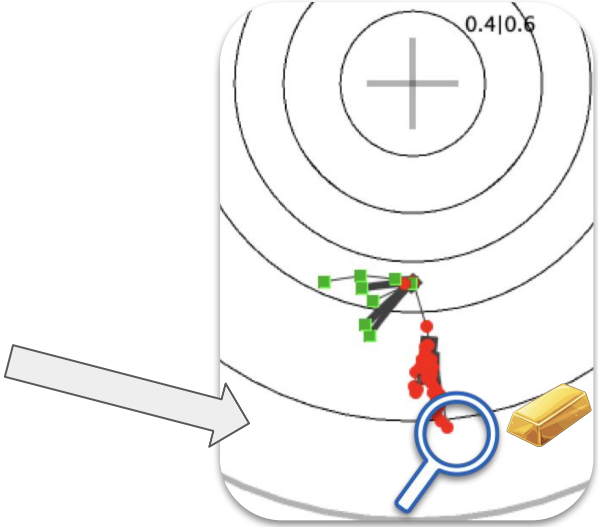
– *Nock et al., "Hyperbolic Embeddings of Supervised Models", NeurIPS'24*

A known numerical issue with Poincaré disk model

- The best parts of a (M)DT embedding are close to the border. In addition to being poorly readable, *numerical approximation issues* can “push” the best confidences to the border, giving a false sense of optimal confidence



- We fix the issue by replacing Riemann summation (at the core of integrals, hence distances), by a *tempered* summation. “Stretches” visualization near border while keeping hyperbolicity



A known numerical issue with Poincaré disk model

- The best parts of a (M)DT embedding are close to the border. In addition to
  being … e best
  confid … e

↪ Includes a generalization of Leibniz-Newton’s fundamental Theorem of calculus
↪ Simple extension of many properties of integration
↪ Gives interesting properties when applied to other models of hyperbolic geometry (e.g. Lorentz’) and beyond, to other integral based “distances” (Bregman divergences, f-divergences, IPMs, etc.). See paper.

- We fix
  (at the
  by a te
  near border while keeping hyperbolicity

Thank You

Richard Nock	Google Research
Ehsan Amid	Google DeepMind
Frank Nielsen	Sony CS Labs, Inc.
Alexander Soen	RIKEN AIP & the Australian National University
Manfred K. Warmuth	Google Research

Google Research