Online Consistency of the Nearest Neighbor Rule

Sanjoy Dasgupta — dasgupta@ucsd.edu

Geelon So — geelon@ucsd.edu

Department of Computer Science and Engineering, UC San Diego

Motivation

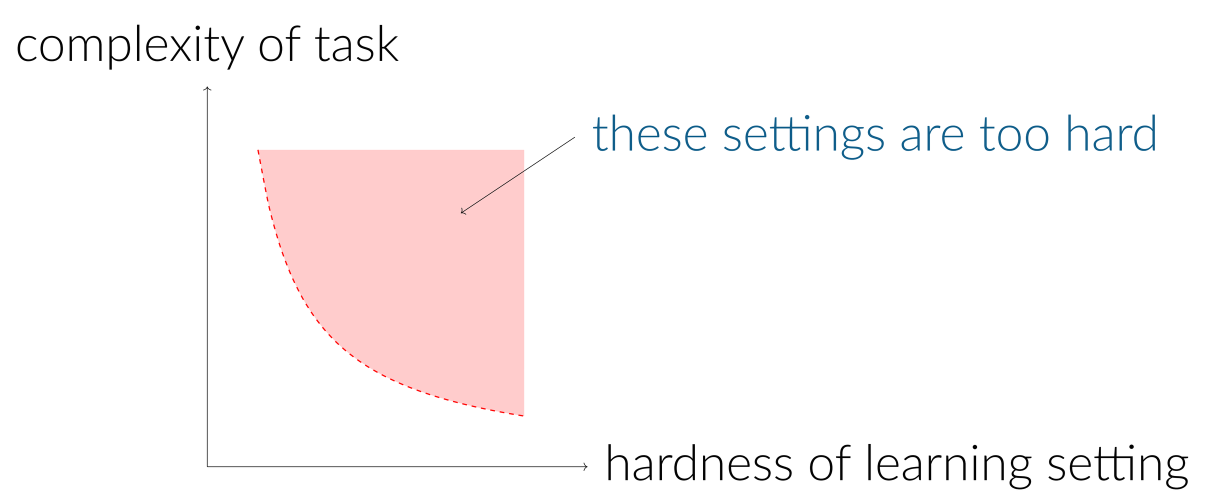
Question: How do we define *realistic* data generating processes?

We don't generally have a good answer to this question, which contributes to the theory-practice gap. We often end up:

- requiring unrealistic/excessively restrictive assumptions, or
- coming to overly pessimistic conclusions.

It is hard to formalize settings that are general enough to capture real-world scenarios but sufficiently constrained to be tractable.

Classical frameworks: statistical (i.i.d.) and worst-case settings



Classical learning theory says that the i.i.d. setting is 'easy' and worst-case is very hard. Very little is known in between, but the aim of smoothed or non-worst-case analysis is to fill in the gap.

Problem setting

Let $\eta : \mathcal{X} \to \mathcal{Y}$ a target function where $\mathcal{X}$ is an instance space and $\mathcal{Y}$ is a finite label space.

Online classification loop (realizable) For $n = 1, 2, \ldots$

- some process generates an instance X_n
- the learner makes a prediction $\hat{Y}_n$
- the ground truth is revealed $Y_n = \eta(X_n)$

Goal of the learner The mistake rate eventually vanishes:

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\left\{\hat{Y}_n \neq Y_n\right\} = 0. \qquad (\star)$$

We say that the learner is *online consistent* on $(\mathbb{X}, \eta)$.

Research goals

- Study the nearest neighbor rule under general settings
- Understand what makes different sequences of data hard

Summary

We study the nearest neighbor rule (1-NN) and show that it learns under settings far broader than previously known.

In the language of non-worst-case analysis, sequences on which the 1-NN rule does not learn are extremely rare—under suitable classes of measures, they have measure zero.

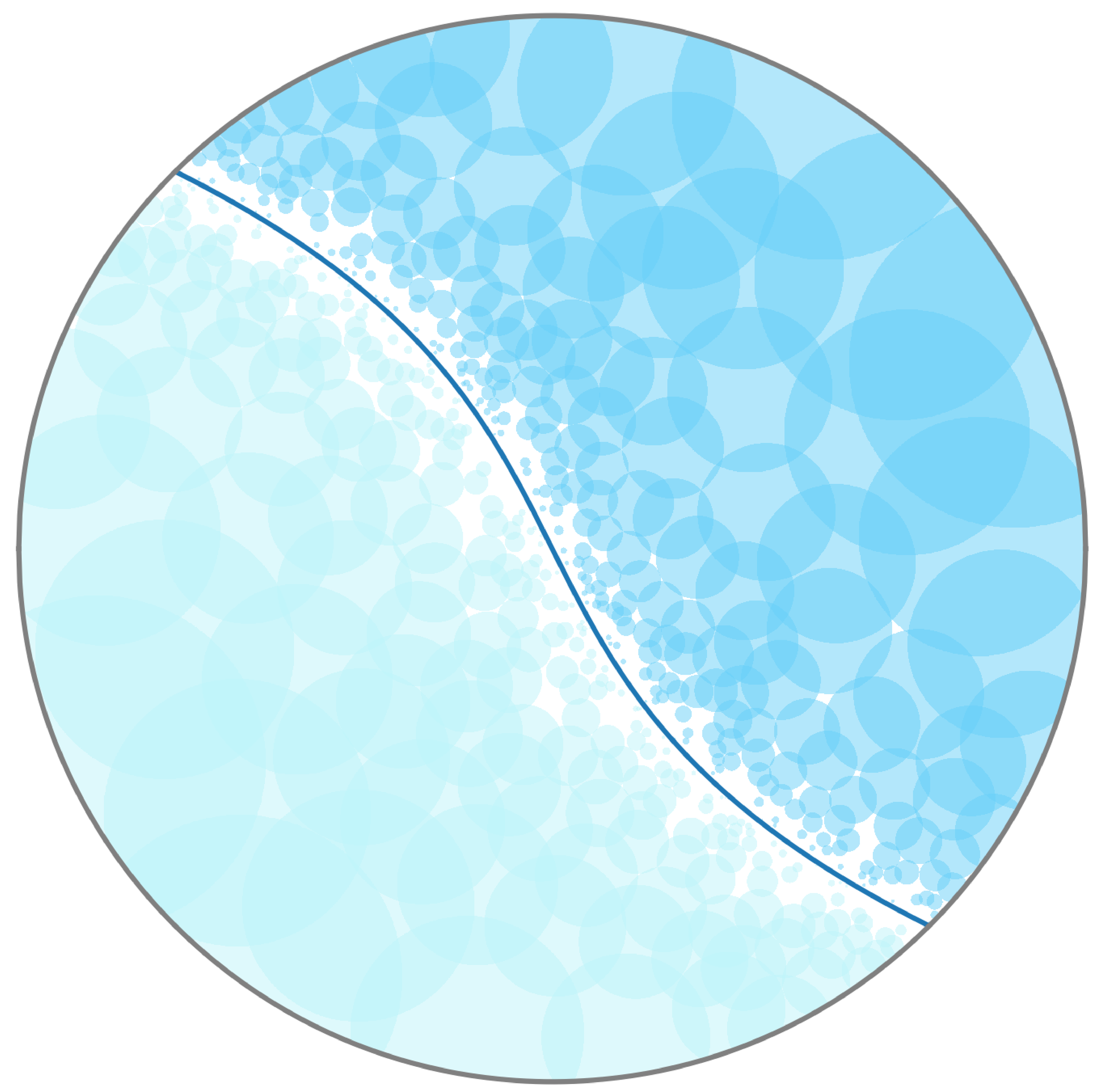


Figure 1. A mutually-labeling ball cover.

The nearest neighbor rule

Let $\mathcal{X}$ have a separable metric ρ and finite Borel measure ν.

- $\mathbb{X} = (X_n)_n$ is an instance sequence generated by a process.
- $\tilde{\mathbb{X}} = (\tilde{X}_n)_n$ is a corresponding *nearest neighbor process*, where:

$$\tilde{X}_n \in \operatorname*{arg\,min}_{x\in\mathbb{X}_{<n}} \rho(X_n, x).$$

- The nearest neighbor rule predicts using the label:

$$\hat{Y}_n = \eta(\tilde{X}_n).$$

Inductive bias: points are surrounded by other points of the same class, provided we zoom in enough.

A class of budgeted processes

Definition. A stochastic process $\mathbb{X}$ is ergodically dominated by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that:

$$\nu(A) < \delta \quad\Longrightarrow\quad \limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left\{X_n \in A\right\} < \varepsilon \qquad \text{a.s.}$$

We say $\mathbb{X}$ is *ergodically continuous* with respect to ν at rate $\varepsilon(\delta)$.

A class of bounded-precision processes

Definition. A stochastic process $\mathbb{X}$ is uniformly dominated by ν if for any $\varepsilon > 0$, there exists $\delta > 0$ such that:

$$\nu(A) < \delta \quad\Longrightarrow\quad \Pr\left(X_n \in A \,\middle|\, \mathbb{X}_{<n}\right) < \varepsilon \qquad \text{a.s.}$$

We say it is *uniformly absolutely continuous* w.r.t. ν at rate $\varepsilon(\delta)$.

Learning in the worst-case setting

Let $(\mathcal{X}, \rho)$ be a totally bounded metric space. Fix $\eta : \mathcal{X} \to \mathcal{Y}$.

Proposition. The nearest neighbor rule is consistent on $(\mathbb{X}, \eta)$ for every $\mathbb{X}$ if and only if the classes are positively separated:

$$\inf_{\eta(x)\neq\eta(x')} \rho(x, x') > c > 0.$$

- Roughly speaking, if and only if the inductive bias is correct.

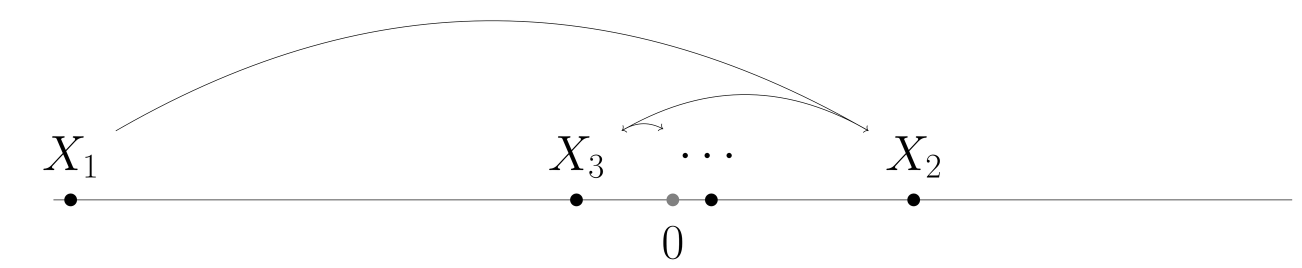


Figure 2. Let $\eta(x) = \mathbf{1}\{x \geq 0\}$. The nearest neighbor rule makes a mistake every single round on the sequence $X_n = (-1/3)^n$.

Learning over nice functions

Theorem. Let $(\mathcal{X}, \rho, \nu)$ be equipped with a separable metric ρ and a finite Borel measure ν. Let η have negligible boundary. When $\mathbb{X}$ is ergodically dominated by ν, then:

$$\underbrace{\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left\{\eta(X_n)\neq\eta(\tilde{X}_n)\right\} = 0}_{\text{the nearest neighbor rule is online consistent for } (\mathbb{X},\eta).} \qquad \text{a.s.}$$

- Here, the inductive bias is correct almost everywhere.

Learning over all functions

Theorem. Additionally, let $(\mathcal{X}, \rho, \nu)$ be upper doubling but η be any measurable function. When $\mathbb{X}$ is uniformly dominated by ν,

$$\underbrace{\limsup_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\left\{\eta(X_n)\neq\eta(\tilde{X}_n)\right\} = 0}_{\text{the nearest neighbor rule is online consistent for } (\mathbb{X},\eta).} \qquad \text{a.s.}$$

- Here, the inductive bias can be correct nowhere.

Geometric notions

- The *margin* of a point x with respect to η is:

$$\mathrm{margin}_\eta(x) = \inf_{\eta(x')\neq\eta(x)} \rho(x, x').$$

- We say x is on the *boundary* $\partial\eta$ if $\mathrm{margin}_\eta(x) = 0$.
- We say that η has *negligible boundary* if $\nu(\partial\eta) = 0$.
- A set $U \subset \mathcal{X}$ is *mutually-labeling* if:

$$\mathrm{diam}(U) < \inf_{x\in U} \mathrm{margin}_\eta(x).$$

  If $x \notin \partial\eta$, then $B(x, \mathrm{margin}_\eta(x)/3)$ is mutually-labeling.
- The space $(\mathcal{X}, \rho, \nu)$ is *upper doubling* with dimension d if:
  - any ball can be covered by 2^d balls with half its radius,
  - r-balls have $O(r^d)$ mass.

Supported by NSF IIS-2211386

UC San Diego