Nearest neighbor classification in metric spaces: universal consistency and rates of convergence

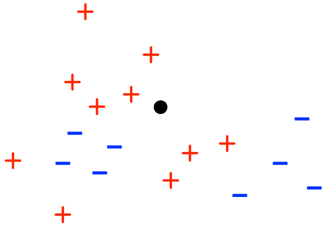
Sanjoy Dasgupta
University of California, San Diego

Nearest neighbor

The primeval approach to classification. Given:

- a *training set* $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ consisting of data points $x_i \in \mathcal{X}$ and their labels $y_i \in \{0, 1\}$
- a *query point* x

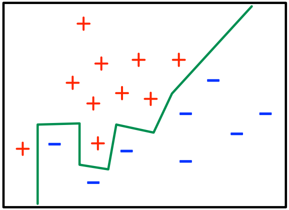
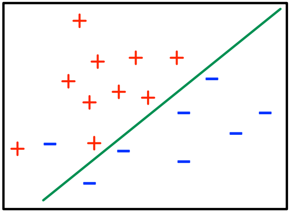
predict the label of x by looking at its nearest neighbor among the x_i.



How accurate is this method? What kinds of data is it well-suited to?

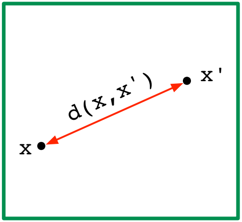
A nonparametric estimator

Contrast with *linear classifiers*, which are also simple and general-purpose.

- **Expressivity**: what kinds of decision boundary can it produce?
- **Consistency**: as the number of points n increases, does the decision boundary converge?
- **Rates of convergence**: how fast does this convergence occur, as a function of n?
- **Style of analysis**.

The data space



Data points lie in a space $\mathcal{X}$ with distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$.

- Most common scenario: $\mathcal{X} = \mathbb{R}^p$ and d is Euclidean distance.
- Our setting: $(\mathcal{X}, d)$ is a *metric space*.
  - ℓ_p distances
  - Metrics obtained from user preferences/feedback
- Also of interest: more general distances.
  - KL divergence
  - Domain-specific dissimilarity measures

Statistical learning theory setup

Training points come from the same source as future query points:

- Underlying measure μ on $\mathcal{X}$ from which all points are generated.
- We call $(\mathcal{X}, d, \mu)$ a *metric measure space*.
- Label of x is a coin flip with bias $\eta(x) = \Pr(Y = 1|X = x)$.

A classifier is a rule $h : \mathcal{X} \to \{0, 1\}$.

- Misclassification rate, or risk: $R(h) = \Pr(h(X) \neq Y)$.
- The *Bayes-optimal classifier*

$$h^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise} \end{cases},$$

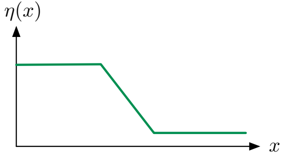
has minimum risk, $R^* = R(h^*) = \mathbb{E}_X \min(\eta(X), 1 - \eta(X))$.

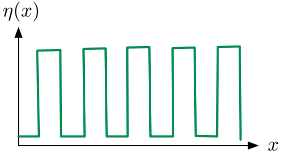
Questions of interest

Let h_n be a classifier based on n labeled data points from the underlying distribution. $R(h_n)$ is a random variable.

- **Consistency**: does $R(h_n)$ converge to R^*?
- **Rates of convergence**: how fast does convergence occur?

The smoothness of $\eta(x) = \Pr(Y = 1|X = x)$ matters:





Questions of interest:

- Consistency without assumptions?
- A suitable smoothness assumption, and rates?
- Rates without assumptions, using distribution-specific quantities?

Talk outline

1. Consistency without assumptions
2. Rates of convergence under smoothness
3. General rates of convergence
4. Open problems

Consistency

Given n data points $(x_1, y_1), \ldots, (x_n, y_n)$, how to answer a query x?

- 1-NN returns the label of the nearest neighbor of x amongst the x_i.
- k-NN returns the majority vote of the k nearest neighbors.
- k_n-NN lets k grow with n.

1-NN and k-NN are not, in general, consistent.

E.g. $\mathcal{X} = \mathbb{R}$ and $\eta(x) \equiv \eta_o < 1/2$. *Every label is a coin flip with bias η_o.*

- Bayes risk is $R^* = \eta_o$ (always predict 0).
- 1-NN risk: what is the probability that two coins of bias η_o disagree?
  $\mathbb{E}R(h_n) = 2\eta_o(1 - \eta_o) > \eta_o$.
- And k-NN has risk $\mathbb{E}R(h_n) = \eta_o + f(k)$.

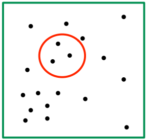
Henceforth h_n denotes the k_n-classifier, where $k_n \to \infty$.

Consistency results under continuity

Assume $\eta(x) = P(Y = 1|X = x)$ is continuous.
Let h_n be the k_n-classifier, with $k_n \uparrow \infty$ and $k_n/n \downarrow 0$.

- Fix and Hodges (1951): Consistent in $\mathbb{R}^p$.
- Cover-Hart (1965, 1967, 1968): Consistent in any metric space.

Proof outline: Let x be a query point and let $x_{(1)}, \ldots, x_{(n)}$ denote the training points ordered by increasing distance from x.



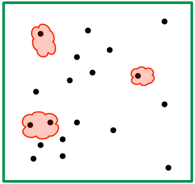
Training points are drawn from μ, so the number of them in any ball B is roughly $n\mu(B)$.

- Therefore $x_{(1)}, \ldots, x_{(k_n)}$ lie in a ball centered at x of probability mass $\approx k_n/n$. Since $k_n/n \downarrow 0$, we have $x_{(1)}, \ldots, x_{(k_n)} \to x$.
- By continuity, $\eta(x_{(1)}), \ldots, \eta(x_{(k_n)}) \to \eta(x)$.
- By law of large numbers, when tossing many coins of bias roughly $\eta(x)$, the fraction of 1s will be approximately $\eta(x)$. Thus the majority vote of their labels will approach $h^*(x)$.

Universal consistency in $\mathbb{R}^p$

Stone (1977): consistency in $\mathbb{R}^p$ assuming only measurability.

Lusin's thm: for any measurable η, for any $\epsilon > 0$, there is a continuous function that differs from it on at most ϵ fraction of points.



Training points in the red region can cause trouble. What fraction of query points have one of these as their nearest neighbor?

Geometric result: pick any set of points in $\mathbb{R}^p$. Then any one point is the NN of at most 5^p other points.

An alternative sufficient condition for arbitrary metric measure spaces $(\mathcal{X}, d, \mu)$: that the fundamental theorem of calculus holds.

Universal consistency in metric spaces

Query x; training points by increasing distance from x are $x_{(1)}, \ldots, x_{(n)}$.

1. Earlier argument: under continuity, $\eta(x_{(1)}), \ldots, \eta(x_{(k_n)}) \to \eta(x)$.
   In this case, the k_n-NN are coins of roughly the same bias as x.
2. It suffices that average$(\eta(x_{(1)}), \ldots, \eta(x_{(k_n)})) \to \eta(x)$.
3. $x_{(1)}, \ldots, x_{(k_n)}$ lie in some ball $B(x, r)$.
   For suitable r, they are random draws from μ restricted to $B(x, r)$.
4. average$(\eta(x_{(1)}), \ldots, \eta(x_{(k_n)}))$ is close to the average η in this ball:

$$\frac{1}{\mu(B(x,r))} \int_{B(x,r)} \eta \, d\mu.$$

5. As n grows, this ball $B(x, r)$ shrinks. Thus it is enough that

$$\lim_{r \downarrow 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} \eta \, d\mu \;=\; \eta(x).$$

   In $\mathbb{R}^p$, this is *Lebesgue's differentiation theorem.*

Universal consistency in metric spaces

Let $(\mathcal{X}, d, \mu)$ be a metric measure space in which the Lebesgue differentiation property holds: for any bounded measurable f,

$$\lim_{r \downarrow 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} f \, d\mu \; = \; f(x)$$

for almost all (μ-a.e.) $x \in \mathcal{X}$.

- If $k_n \to \infty$ and $k_n/n \to 0$, then $R_n \to R^*$ in probability.
- If in addition $k_n/\log n \to \infty$, then $R_n \to R^*$ almost surely.

Examples of such spaces: finite-dimensional normed spaces; doubling metric measure spaces.

Talk outline

1. Consistency without assumptions
2. Rates of convergence under smoothness
3. General rates of convergence
4. Open problems

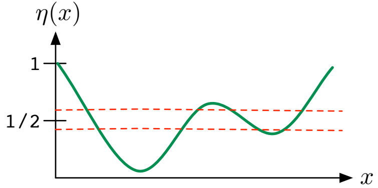
Smoothness and margin conditions

- The usual smoothness condition in $\mathbb{R}^p$: η is α-Holder continuous if for some constant L, for all x, x',

$$|\eta(x) - \eta(x')| \leq L\|x - x'\|^\alpha.$$

- Mammen-Tsybakov β-margin condition: For some constant C, for any t, we have $\mu(\{x : |\eta(x) - 1/2| \leq t\}) \leq Ct^\beta$.

Width-*t margin* around decision boundary

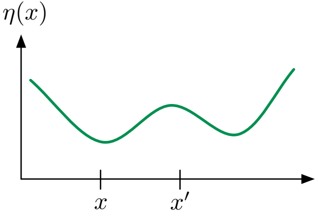


- Audibert-Tsybakov: Suppose these two conditions hold, and that μ is supported on a *regular* set with $0 < \mu_{min} < \mu < \mu_{max}$. Then $\mathbb{E}R_n - R^*$ is $\Omega(n^{-\alpha(\beta+1)/(2\alpha+p)})$.

Under these conditions, for suitable (k_n), this rate is achieved by k_n-NN.

A better smoothness condition for NN

How much does η change over an interval?



- The usual notions relate this to $|x - x'|$.
- For NN: more sensible to relate to $\mu([x, x'])$.

We will say η **is α-smooth in metric measure space** $(\mathcal{X}, d, \mu)$ if for some constant L, for all $x \in \mathcal{X}$ and $r > 0$,

$$|\eta(x) - \eta(B(x,r))| \;\leq\; L\,\mu(B(x,r))^{\alpha},$$

where $\eta(B) =$ average η in ball $B = \frac{1}{\mu(B)} \int_B \eta \; d\mu$.

η is α-Holder continuous in $\mathbb{R}^p$, μ bounded below $\Rightarrow$ η is (α/p)-smooth.

Rates of convergence under smoothness

Let $h_{n,k}$ denote the k-NN classifier based on n training points.
Let h^* be the Bayes-optimal classifier.

Suppose η is α-smooth in $(\mathcal{X}, d, \mu)$. Then for any n, k,

1. For any $\delta > 0$, with probability at least $1 - \delta$ over the training set,
   $\Pr_X(h_{n,k}(X) \neq h^*(X)) \ \leq \ \delta + \mu(\{x : |\eta(x) - \frac{1}{2}| \leq C_1 \sqrt{\frac{1}{k} \ln \frac{1}{\delta}}\})$
   under the choice $k \propto n^{2\alpha/(2\alpha+1)}$.
2. $\mathbb{E}_n \Pr_X(h_{n,k}(X) \neq h^*(X)) \ \geq \ C_2\, \mu(\{x : |\eta(x) - \frac{1}{2}| \leq C_3 \sqrt{\frac{1}{k}}\})$.

These upper and lower bounds are qualitatively similar for *all* smooth conditional probability functions:

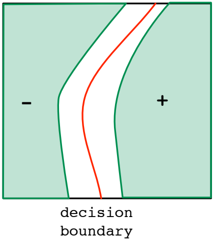
> *the probability mass of the width-$\frac{1}{\sqrt{k}}$ margin around the decision boundary.*

Talk outline

1. Consistency without assumptions
2. Rates of convergence under smoothness
3. General rates of convergence
4. Open problems

General rates of convergence

For sample size n, can identify positive and negative regions that will reliably be classified:



- *Probability-radius*: Grow a ball around x until probability mass $\geq p$:

$$r_p(x) = \inf\{r : \mu(B(x,r)) \geq p\}.$$

Probability-radius of interest: $p = k/n$.

- Reliable positive region:

$$\mathcal{X}^+_{p,\Delta} = \{x : \eta(B(x,r)) \geq \frac{1}{2} + \Delta \text{ for all } r \leq r_p(x)\}$$

where $\Delta \approx 1/\sqrt{k}$. Likewise negative region, $\mathcal{X}^-_{p,\Delta}$.

- Effective boundary: $\partial_{p,\Delta} = \mathcal{X} \setminus (\mathcal{X}^+_{p,\Delta} \cup \mathcal{X}^-_{p,\Delta})$.

Roughly, $\Pr_X(h_{n,k}(X) \neq h^*(X)) \leq \mu(\partial_{p,\Delta})$.

Open problems

1. Necessary and sufficient conditions for universal consistency in metric measure spaces.
2. Consistency in more general topological spaces.
3. Extension to countably infinite label spaces.
4. Applications of convergence rates: active learning, domain adaptation, ...

Thanks

To my co-author Kamalika Chaudhuri and to the National Science Foundation for support under grant IIS-1162581.