Distribution-specific analysis of nearest neighbor search and classification

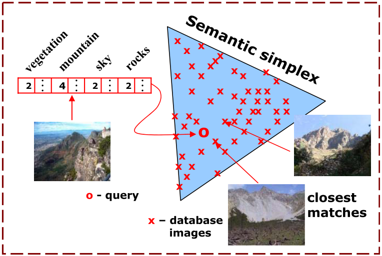
Sanjoy Dasgupta
University of California, San Diego

Nearest neighbor

The primeval approach to information retrieval and classification.

Example: image retrieval and classification.
Given a query image, find similar images in a database using NN search.

E.g. Fixed-dimensional "semantic representation":

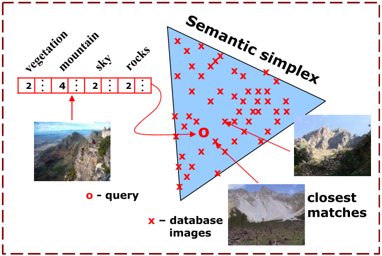


Nearest neighbor

The primeval approach to information retrieval and classification.

Example: image retrieval and classification.
Given a query image, find similar images in a database using NN search.

E.g. Fixed-dimensional "semantic representation":



Basic questions:

- Statistical: error analysis of NN classification
- Algorithmic: finding the NN quickly

Rate of convergence of NN classification

The data distribution:

- Data points X are drawn from a distribution μ on $\mathbb{R}^p$
- Labels $Y \in \{0, 1\}$ follow $\Pr(Y = 1 | X = x) = \eta(x)$.

Rate of convergence of NN classification

The data distribution:

- Data points X are drawn from a distribution μ on $\mathbb{R}^p$
- Labels $Y \in \{0, 1\}$ follow $\Pr(Y = 1|X = x) = \eta(x)$.

Classical theory for NN (or k-NN) classifier based on n data points:

- Can we give a non-asymptotic error bound depending only on n, p?

Rate of convergence of NN classification

The data distribution:

- Data points X are drawn from a distribution μ on $\mathbb{R}^p$
- Labels $Y \in \{0, 1\}$ follow $\Pr(Y = 1 | X = x) = \eta(x)$.

Classical theory for NN (or k-NN) classifier based on n data points:

- Can we give a non-asymptotic error bound depending only on n, p?
  No.

Rate of convergence of NN classification

The data distribution:

- Data points X are drawn from a distribution μ on $\mathbb{R}^p$
- Labels $Y \in \{0, 1\}$ follow $\Pr(Y = 1|X = x) = \eta(x)$.

Classical theory for NN (or k-NN) classifier based on n data points:

- Can we give a non-asymptotic error bound depending only on n, p? No.
- Smoothness assumption: η is α-Holder continuous:

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

Rate of convergence of NN classification

The data distribution:

- Data points X are drawn from a distribution μ on $\mathbb{R}^p$
- Labels $Y \in \{0, 1\}$ follow $\Pr(Y = 1|X = x) = \eta(x)$.

Classical theory for NN (or k-NN) classifier based on n data points:

- Can we give a non-asymptotic error bound depending only on n, p?
  No.
- Smoothness assumption: η is α-Holder continuous:

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

  Then: error bound $O(n^{-\alpha/(2\alpha+p)})$.

Rate of convergence of NN classification

The data distribution:

- Data points X are drawn from a distribution μ on $\mathbb{R}^p$
- Labels $Y \in \{0, 1\}$ follow $\Pr(Y = 1|X = x) = \eta(x)$.

Classical theory for NN (or k-NN) classifier based on n data points:

- Can we give a non-asymptotic error bound depending only on n, p? No.
- Smoothness assumption: η is α-Holder continuous:

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

  Then: error bound $O(n^{-\alpha/(2\alpha+p)})$.
- This is "optimal".

Rate of convergence of NN classification

The data distribution:

- Data points X are drawn from a distribution μ on $\mathbb{R}^p$
- Labels $Y \in \{0, 1\}$ follow $\Pr(Y = 1 | X = x) = \eta(x)$.

Classical theory for NN (or k-NN) classifier based on n data points:

- Can we give a non-asymptotic error bound depending only on n, p? No.
- Smoothness assumption: η is α-Holder continuous:

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

  Then: error bound $O(n^{-\alpha/(2\alpha+p)})$.
- This is "optimal".
  There exists a distribution with parameter α for which this bound is achieved.

Goals

What we need for nonparametric estimators like NN:

1. Bounds that hold without any assumptions.
   Use these to determine parameters that truly govern the difficulty of the problem.
2. How do we know when the bounds are tight enough?
   When the lower and upper bounds are comparable **for every instance**.

The complexity of nearest neighbor search

Given a data set of n points in $\mathbb{R}^p$, build a data structure for efficiently answering subsequent nearest neighbor queries q.

- Data structure should take space $O(n)$
- Query time should be $o(n)$

The complexity of nearest neighbor search

Given a data set of n points in $\mathbb{R}^p$, build a data structure for efficiently answering subsequent nearest neighbor queries q.

- Data structure should take space $O(n)$
- Query time should be $o(n)$

Troubling example: exponential dependence on dimension?

For any $0 < \epsilon < 1$,

- Pick $2^{O(\epsilon^2 p)}$ points uniformly from the unit sphere in $\mathbb{R}^p$
- With high probability, all interpoint distances are $(1 \pm \epsilon)\sqrt{2}$

Approximate nearest neighbor

For data set $S \subset \mathbb{R}^p$ and query q, a c-approximate nearest neighbor is any $x \in S$ such that

$$\|x - q\| \leq c \cdot \min_{z \in S} \|z - q\|.$$

Approximate nearest neighbor

For data set $S \subset \mathbb{R}^p$ and query q, a c-approximate nearest neighbor is any $x \in S$ such that

$$\|x - q\| \leq c \cdot \min_{z \in S} \|z - q\|.$$

Locality-sensitive hashing (Indyk, Motwani, Andoni):

- Data structure size n^{1+1/c^2}
- Query time n^{1/c^2}

Approximate nearest neighbor

For data set $S \subset \mathbb{R}^p$ and query q, a c-approximate nearest neighbor is any $x \in S$ such that

$$\|x - q\| \leq c \cdot \min_{z \in S} \|z - q\|.$$

Locality-sensitive hashing (Indyk, Motwani, Andoni):

- Data structure size n^{1+1/c^2}
- Query time n^{1/c^2}

Is "c" a good measure of the hardness of the problem?

Approximate nearest neighbor

The MNIST data set of handwritten digits:

What % of c-approximate nearest neighbors have the wrong label?

Approximate nearest neighbor

The MNIST data set of handwritten digits:

What % of c-approximate nearest neighbors have the wrong label?

c	1.0	1.2	1.4	1.6	1.8	2.0
Error rate (%)	3.1	9.0	18.4	29.3	40.7	51.4

Goals

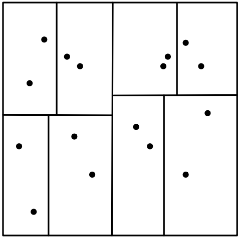
What we need for nonparametric estimators like NN:

1. Bounds that hold without any assumptions.
   Use these to determine parameters that truly govern the difficulty of the problem.
2. How do we know when the bounds are tight enough?
   When the lower and upper bounds are comparable **for every instance**.

Talk outline

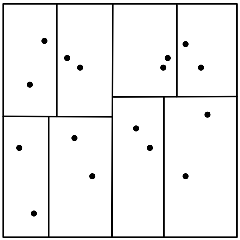
1. Complexity of NN search
2. Rates of convergence for NN classification

The k-d tree: a hierarchical partition of $\mathbb{R}^p$



Defeatist search: return NN in query point's leaf node.

The k-d tree: a hierarchical partition of $\mathbb{R}^p$



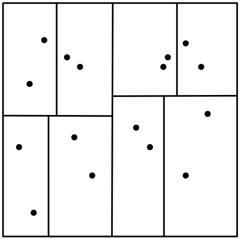
Defeatist search: return NN in query point's leaf node.

Problem: This might fail to return the true NN.

Heuristics for reducing failure probability in high dimension:

- Random split directions (Liu, Moore, Gray, and Kang)
- Overlapping cells (Maneewongvatana and Mount; Liu et al)

The k-d tree: a hierarchical partition of $\mathbb{R}^p$



Defeatist search: return NN in query point's leaf node.

Problem: This might fail to return the true NN.

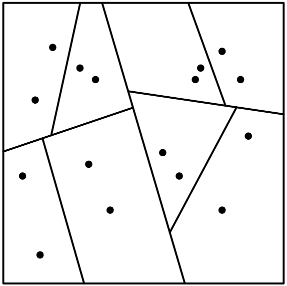
Heuristics for reducing failure probability in high dimension:

- Random split directions (Liu, Moore, Gray, and Kang)
- Overlapping cells (Maneewongvatana and Mount; Liu et al)

Popular option: forests of randomized trees (e.g. FLANN)

Heuristic 1: Random split directions

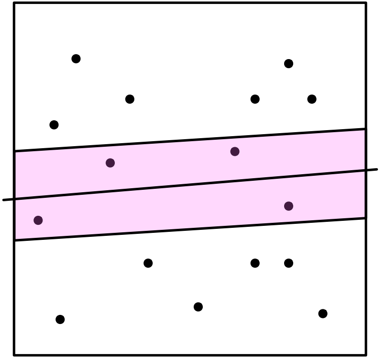
In each cell of the tree, pick split direction uniformly at random from the unit sphere in $\mathbb{R}^p$



Perturbed split: after projection, pick $\beta \in_R [1/4, 3/4]$ and split at the β-fractile point.

Heuristic 2: Overlapping cells

Overlapping split points: $1/2 - \alpha$ fractile and $1/2 + \alpha$ fractile

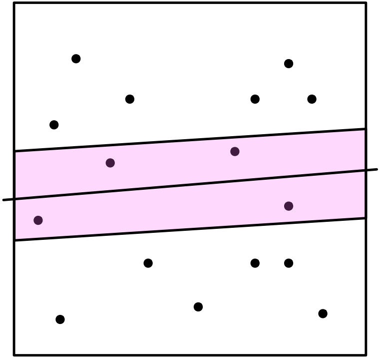


Procedure:

- Route data (to multiple leaves) using overlapping splits
- Route query (to single leaf) using median split

Heuristic 2: Overlapping cells

Overlapping split points: $1/2 - \alpha$ fractile and $1/2 + \alpha$ fractile

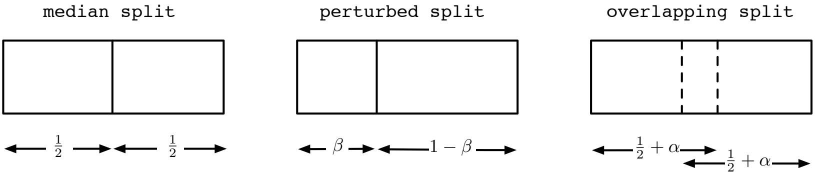


Procedure:

- Route data (to multiple leaves) using overlapping splits
- Route query (to single leaf) using median split

Spill tree has size $n^{1/(1-\lg(1+2\alpha))}$: e.g. $n^{1.159}$ for $\alpha = 0.05$.

Two randomized partition trees



	Routing data	Routing queries
k-d tree	Median split	Median split
Random projection tree	Perturbed split	Perturbed split
Spill tree	Overlapping split	Median split

Failure probability

Pick any data set $x_1, \ldots, x_n$ and any query q.

- Let $x_{(1)}, \ldots, x_{(n)}$ be the ordering of data by distance from q.
- Probability of not returning the NN depends directly on

$$\Phi(q, \{x_1, \ldots, x_n\}) = \frac{1}{n} \sum_{i=2}^{n} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

(This probability is over the randomization in tree construction.)

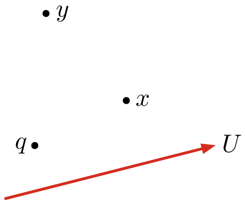
- Spill tree: failure probability $\propto \Phi$
- RP tree: failure probability $\propto \Phi \log 1/\Phi$

Random projection of three points

Let $q \in \mathbb{R}^p$ be the query, x its nearest neighbor and y some other point:

$$\|q - x\| < \|q - y\|.$$

Bad event: when the data is projected onto a random direction U, point y falls between q and x.



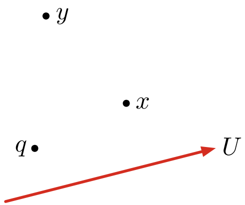
What is the probability of this?

Random projection of three points

Let $q \in \mathbb{R}^p$ be the query, x its nearest neighbor and y some other point:

$$\|q - x\| < \|q - y\|.$$

Bad event: when the data is projected onto a random direction U, point y falls between q and x.

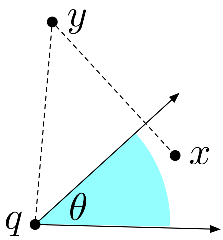


What is the probability of this?

This is a 2-d problem, in the plane defined by q, x, y.

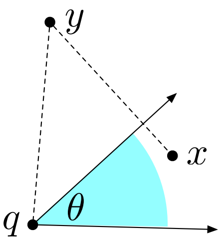
- Only care about projection of U on this plane
- Projection of U is a random direction in this plane

Random projection of three points



Probability that U falls in this bad region is $\theta/2\pi$.

Random projection of three points



Probability that U falls in this bad region is $\theta/2\pi$.

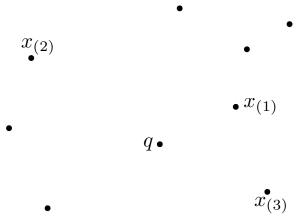
Lemma

Pick any three points $q, x, y \in \mathbb{R}^p$ such that $\|q - x\| < \|q - y\|$. Pick U uniformly at random from the unit sphere S^{p-1}. Then

$$Pr(y \cdot U \text{ falls between } q \cdot U \text{ and } x \cdot U) \leq \frac{1}{2} \frac{\|q - x\|}{\|q - y\|}.$$

(Tight within a constant unless the points are almost-collinear)

Random projection of a set of points



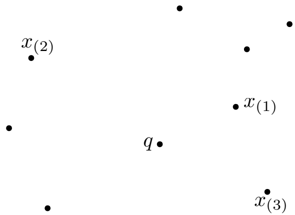
Lemma

Pick any $x_1, \ldots, x_n$ and any query q. Pick $U \in_R S^{p-1}$ and project all points onto direction U. Expected fraction of projected x_i falling between q and $x_{(1)}$ is at most

$$\frac{1}{2n} \sum_{i=2}^{n} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|} \;=\; \frac{1}{2}\Phi$$

Proof: Probability that $x_{(i)}$ falls between q and $x_{(1)}$ is at most $\frac{1}{2}\frac{\|q-x_{(1)}\|}{\|q-x_{(i)}\|}$. Now use linearity of expectation.

Random projection of a set of points



Lemma

Pick any $x_1, \ldots, x_n$ and any query q. Pick $U \in_R S^{p-1}$ and project all points onto direction U. Expected fraction of projected x_i falling between q and $x_{(1)}$ is at most

$$\frac{1}{2n} \sum_{i=2}^{n} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|} = \frac{1}{2}\Phi$$

Proof: Probability that $x_{(i)}$ falls between q and $x_{(1)}$ is at most $\frac{1}{2}\frac{\|q-x_{(1)}\|}{\|q-x_{(i)}\|}$. Now use linearity of expectation.

Bad event: this fraction is $> \alpha n$. Happens with probability $\leq \Phi/2\alpha$.

Failure probability of NN search

Fix any data points $x_1, \ldots, x_n$ and query q. For $m \leq n$, define

$$\Phi_m(q, \{x_1, \ldots, x_n\}) = \frac{1}{m} \sum_{i=2}^{m} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

Failure probability of NN search

Fix any data points $x_1, \ldots, x_n$ and query q. For $m \leq n$, define

$$\Phi_m(q, \{x_1, \ldots, x_n\}) = \frac{1}{m} \sum_{i=2}^{m} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

Theorem

Suppose a randomized spill tree is built for data set $x_1, \ldots, x_n$ with leaf nodes of size n_o. For any query q, the probability that the NN query does not return $x_{(1)}$ is at most

$$\frac{1}{2\alpha} \sum_{i=0}^{\ell} \Phi_{\beta^i n}(q, \{x_1, \ldots, x_n\})$$

where $\beta = 1/2 + \alpha$ and $\ell = \log_{1/\beta}(n/n_o)$ is the tree's depth.

- RP tree: same result, with $\beta = 3/4$ and $\Phi \to \Phi \ln(2e/\Phi)$
- Extension to k nearest neighbors is immediate

Bounding Φ in cases of interest

Need to bound

$$\Phi_m(q, \{x_1, \ldots, x_n\}) = \frac{1}{m} \sum_{i=2}^{m} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

What structural assumptions on the data might be suitable?

Bounding Φ in cases of interest

Need to bound

$$\Phi_m(q, \{x_1, \ldots, x_n\}) = \frac{1}{m} \sum_{i=2}^{m} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

What structural assumptions on the data might be suitable?

Set $S \subset \mathbb{R}^p$ has *doubling dimension* k if for any (Euclidean) ball B, the subset $S \cap B$ can be covered by 2^k balls of half the radius.

Bounding Φ in cases of interest

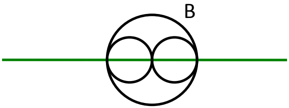
Need to bound

$$\Phi_m(q, \{x_1, \ldots, x_n\}) = \frac{1}{m} \sum_{i=2}^{m} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

What structural assumptions on the data might be suitable?

Set $S \subset \mathbb{R}^p$ has *doubling dimension* k if for any (Euclidean) ball B, the subset $S \cap B$ can be covered by 2^k balls of half the radius.

Example: $S =$ line has doubling dimension 1.



Bounding Φ in cases of interest

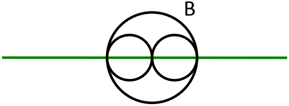
Need to bound

$$\Phi_m(q, \{x_1, \ldots, x_n\}) = \frac{1}{m} \sum_{i=2}^{m} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

What structural assumptions on the data might be suitable?

Set $S \subset \mathbb{R}^p$ has *doubling dimension* k if for any (Euclidean) ball B, the subset $S \cap B$ can be covered by 2^k balls of half the radius.

Example: $S =$ line has doubling dimension 1.



Also generalizes k-dimensional flat, k-dimensional Riemannian submanifold of bounded curvature, k-sparse sets.

NN search in spaces of bounded doubling dimension

Need to bound

$$\Phi_m(q, \{x_1, \ldots, x_n\}) = \frac{1}{m} \sum_{i=2}^{m} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

Suppose:

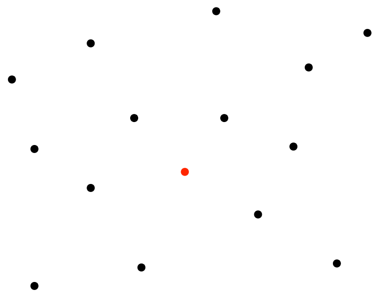
- Pick any $n+1$ points in $\mathbb{R}^p$ with doubling dimension k
- Randomly pick one of them as q; the rest are $x_1, \ldots, x_n$

Then $\mathbb{E}\Phi_m \leq 1/m^{1/k}$.

For constant expected failure probability, use spill tree with leaf size $n_o = O(k^k)$, and query time $O(n_o + \log n)$.

How does doubling dimension help?

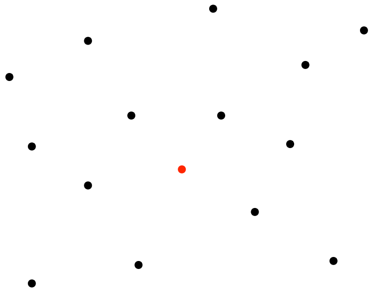
Pick any n points in $\mathbb{R}^p$. Pick one of these points, x. At most how many of the remaining points can have x as its nearest neighbor?



How does doubling dimension help?

Pick any n points in $\mathbb{R}^p$. Pick one of these points, x. At most how many of the remaining points can have x as its nearest neighbor?

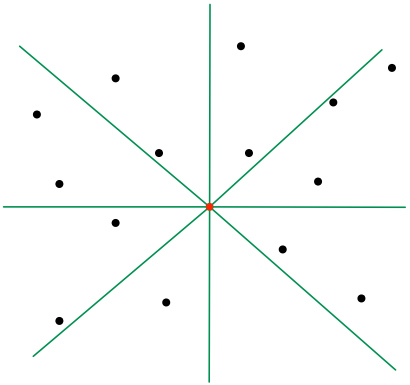
At most c^p, for some constant c [Stone].



How does doubling dimension help?

Pick any n points in $\mathbb{R}^p$. Pick one of these points, x. At most how many of the remaining points can have x as its nearest neighbor?

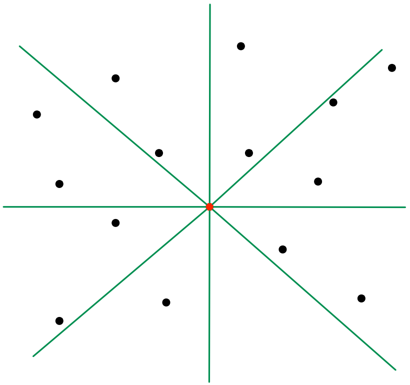
At most c^p, for some constant c [Stone].



How does doubling dimension help?

Pick any n points in $\mathbb{R}^p$. Pick one of these points, x. At most how many of the remaining points can have x as its nearest neighbor?

At most c^p, for some constant c [Stone].



Can (almost) replace p by the doubling dimension [Clarkson].

Open problems

1. Formalizing helpful structure in data.
   What are other types of structure in data for which

   $$\Phi(q, \{x_1, \ldots, x_n\}) = \frac{1}{n} \sum_{i=2}^{n} \frac{\|q - x_{(1)}\|}{\|q - x_{(i)}\|}$$

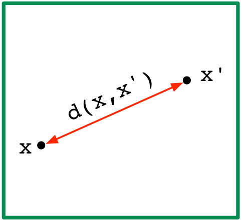
   can be bounded?

2. Empirical study of Φ.
   Is Φ a good predictor of which NN search problems are harder than others?

Talk outline

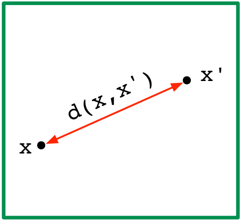
1. Complexity of NN search
2. Rates of convergence for NN classification

Nearest neighbor classification



Data points lie in a metric space $(\mathcal{X}, d)$.

Nearest neighbor classification



Data points lie in a metric space $(\mathcal{X}, d)$.

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.
- Often let k grow with n.

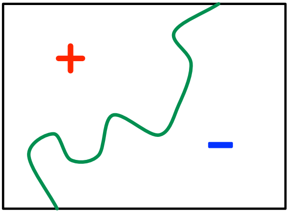
Statistical learning theory setup

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

- Underlying measure μ on $\mathcal{X}$ from which all points are generated.
- Label Y of X follows distribution $\eta(x) = \Pr(Y = 1 | X = x)$.
- The *Bayes-optimal classifier*

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

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



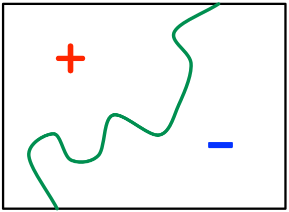
Statistical learning theory setup

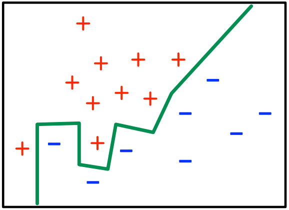
Training points come from the same source as future query points:

- Underlying measure μ on $\mathcal{X}$ from which all points are generated.
- Label Y of X follows distribution $\eta(x) = \Pr(Y = 1 | X = x)$.
- The *Bayes-optimal classifier*

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

  has the minimum possible error, $R^* = \mathbb{E}_X \min(\eta(X), 1 - \eta(X))$.
- **Let $h_{n,k}$ be the k-NN classifier based on n labeled data points.**





Questions of interest

Let $h_{n,k}$ be the k-NN classifier based on n labeled data points.

1. Bounding the error of $h_{n,k}$.

Questions of interest

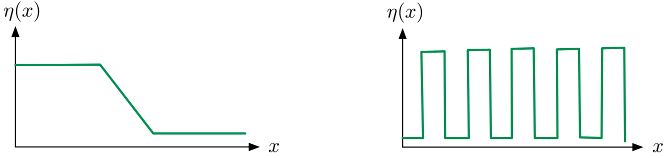
Let $h_{n,k}$ be the k-NN classifier based on n labeled data points.

1. Bounding the error of $h_{n,k}$.
   Assumption-free bounds on $\Pr(h_{n,k}(X) \neq h^*(X))$.

Questions of interest

Let $h_{n,k}$ be the k-NN classifier based on n labeled data points.

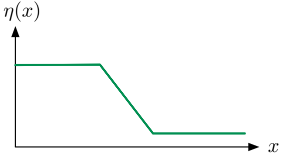
1. Bounding the error of $h_{n,k}$.
   Assumption-free bounds on $\Pr(h_{n,k}(X) \neq h^*(X))$.
2. Smoothness.
   The smoothness of $\eta(x) = \Pr(Y = 1 | X = x)$ matters:



Questions of interest

Let $h_{n,k}$ be the k-NN classifier based on n labeled data points.

1. Bounding the error of $h_{n,k}$.
   Assumption-free bounds on $\Pr(h_{n,k}(X) \neq h^*(X))$.
2. Smoothness.
   The smoothness of $\eta(x) = \Pr(Y = 1|X = x)$ matters:

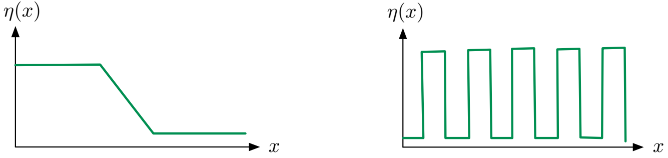


- A notion of smoothness tailor-made for NN.

Questions of interest

Let $h_{n,k}$ be the k-NN classifier based on n labeled data points.

1. Bounding the error of $h_{n,k}$.
   Assumption-free bounds on $\Pr(h_{n,k}(X) \neq h^*(X))$.
2. Smoothness.
   The smoothness of $\eta(x) = \Pr(Y = 1|X = x)$ matters:

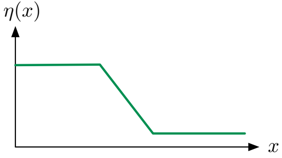


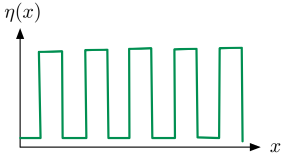
- A notion of smoothness tailor-made for NN.
- Upper and lower bounds that are qualitatively similar for **all** distributions in the same smoothness class.

Questions of interest

Let $h_{n,k}$ be the k-NN classifier based on n labeled data points.

1. Bounding the error of $h_{n,k}$.
   Assumption-free bounds on $\Pr(h_{n,k}(X) \neq h^*(X))$.
2. Smoothness.
   The smoothness of $\eta(x) = \Pr(Y = 1|X = x)$ matters:



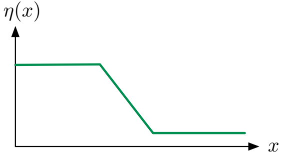


   - A notion of smoothness tailor-made for NN.
   - Upper and lower bounds that are qualitatively similar for **all** distributions in the same smoothness class.
3. Consistency of NN
   Earlier work: Universal consistency in $\mathbb{R}^p$ [Stone]

Questions of interest

Let $h_{n,k}$ be the k-NN classifier based on n labeled data points.

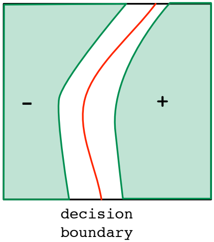
1. Bounding the error of $h_{n,k}$.
   Assumption-free bounds on $\Pr(h_{n,k}(X) \neq h^*(X))$.
2. Smoothness.
   The smoothness of $\eta(x) = \Pr(Y = 1|X = x)$ matters:



   - A notion of smoothness tailor-made for NN.
   - Upper and lower bounds that are qualitatively similar for **all** distributions in the same smoothness class.
3. Consistency of NN
   Earlier work: Universal consistency in $\mathbb{R}^p$ [Stone]
   Now: Universal consistency in a richer family of metric spaces.

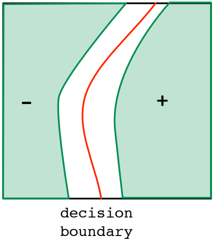
General rates of convergence

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



General rates of convergence

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



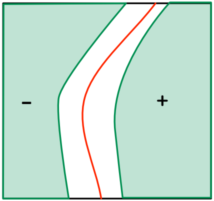
- For any ball B, let $\mu(B)$ be its probability mass and $\eta(B)$ its average η-value, i.e. $\eta(B) = \frac{1}{\mu(B)} \int_B \eta \, d\mu$.
- *Probability-radius*: Grow a ball around x until probability mass $\geq p$:

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

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

General rates of convergence

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



decision boundary

- For any ball B, let $\mu(B)$ be its probability mass and $\eta(B)$ its average η-value, i.e. $\eta(B) = \frac{1}{\mu(B)} \int_B \eta \, d\mu$.
- *Probability-radius*: Grow a ball around x until probability mass $\geq p$:

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

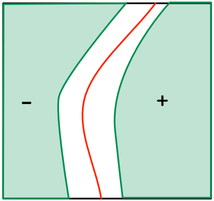
  Probability-radius of interest: $p = k/n$.
- Reliable positive region:

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

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

General rates of convergence

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



- For any ball B, let $\mu(B)$ be its probability mass and $\eta(B)$ its average η-value, i.e. $\eta(B) = \frac{1}{\mu(B)} \int_B \eta \, d\mu$.
- *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:

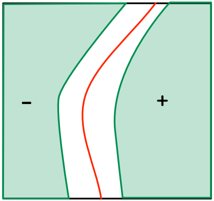
$$\mathcal{X}^+_{p,\Delta} = \{x : \eta(B(x,r)) \geq \frac{1}{2} + \Delta \text{ for all } r \leq r_p(x)\}$$

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

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

General rates of convergence

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



- For any ball B, let $\mu(B)$ be its probability mass and $\eta(B)$ its average η-value, i.e. $\eta(B) = \frac{1}{\mu(B)} \int_B \eta \, d\mu$.
- *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, $\Pr_X(h_{n,k}(X) \neq h^*(X)) \leq \mu(\partial_{p,\Delta})$.

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',

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

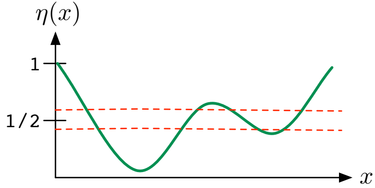
Smoothness and margin conditions

- 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 $\mu\left(\{x : |\eta(x) - 1/2| \leq t\}\right) \leq Ct^\beta$.

Width-t *margin* around decision boundary



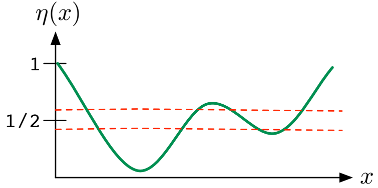
Smoothness and margin conditions

- 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 $\Omega(n^{-\alpha(\beta+1)/(2\alpha+p)})$.

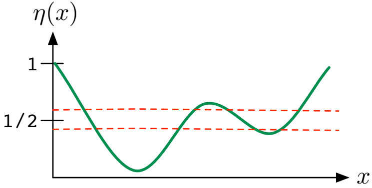
Smoothness and margin conditions

- 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 $\mu\left(\{x : |\eta(x) - 1/2| \leq t\}\right) \leq Ct^\beta$.

Width-*t margin* around decision boundary

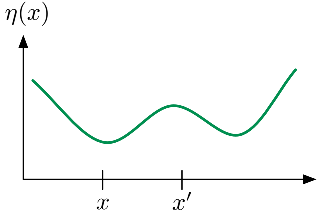


- Audibert-Tsybakov: Suppose these two conditions hold, and that μ is supported on a *regular* set with $0 < \mu_{min} < \mu < \mu_{max}$. Then $\mathbb{E}R_n - R^*$ is $\Omega(n^{-\alpha(\beta+1)/(2\alpha+p)})$.

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

A better smoothness condition for NN

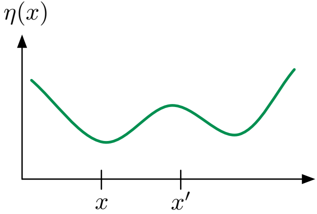
How much does η change over an interval?



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

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 $\mu([x, x'])$.

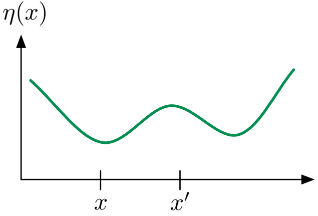
We will say η **is α-smooth in metric measure space** $(\mathcal{X}, d, \mu)$ if for some constant L, for all $x \in \mathcal{X}$ and $r > 0$,

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

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

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,
$$\Pr_X(h_{n,k}(X) \neq h^*(X)) \;\leq\; \delta + \mu(\{x : |\eta(x) - \tfrac{1}{2}| \leq C_1 \sqrt{\tfrac{1}{k} \ln \tfrac{1}{\delta}}\})$$
under the choice $k \propto n^{2\alpha/(2\alpha+1)}$.
2. $$\mathbb{E}_n \Pr_X(h_{n,k}(X) \neq h^*(X)) \;\geq\; C_2\, \mu(\{x : |\eta(x) - \tfrac{1}{2}| \leq C_3 \sqrt{\tfrac{1}{k}}\}).$$

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,
$$\mathrm{Pr}_X(h_{n,k}(X) \neq h^*(X)) \;\leq\; \delta + \mu(\{x : |\eta(x) - \tfrac{1}{2}| \leq C_1 \sqrt{\tfrac{1}{k} \ln \tfrac{1}{\delta}}\})$$
under the choice $k \propto n^{2\alpha/(2\alpha+1)}$.
2. $$\mathbb{E}_n \Pr_X(h_{n,k}(X) \neq h^*(X)) \;\geq\; C_2\, \mu(\{x : |\eta(x) - \tfrac{1}{2}| \leq C_3 \sqrt{\tfrac{1}{k}}\}).$$

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.*

Universal consistency in metric spaces

- Let R_n be error of k-NN classifier and R^* the Bayes-optimal error.
- Universal consistency: $R_n \to R^*$ (for a suitable schedule of k), no matter what the distribution.
- Stone (1977): universal consistency in $\mathbb{R}^p$.

Universal consistency in metric spaces

- Let R_n be error of k-NN classifier and R^* the Bayes-optimal error.
- Universal consistency: $R_n \to R^*$ (for a suitable schedule of k), no matter what the distribution.
- Stone (1977): universal consistency in $\mathbb{R}^p$.

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.

Universal consistency in metric spaces

- Let R_n be error of k-NN classifier and R^* the Bayes-optimal error.
- Universal consistency: $R_n \to R^*$ (for a suitable schedule of k), no matter what the distribution.
- Stone (1977): universal consistency in $\mathbb{R}^p$.

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.

Open problems

1. Are there metric spaces in which k-NN fails to be consistent?

Open problems

1. Are there metric spaces in which k-NN fails to be consistent?
2. Consistency in more general distance spaces.