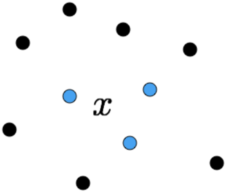
Understanding Thy Neighbors:
Practical Perspectives from Modern Analysis



Sanjoy Dasgupta
CSE, UCalifornia, San Diego

Samory Kpotufe
ORFE, Princeton University

k-Nearest Neighbor Approach:

Use the k closest datapoints to x *to infer* something about x.

Ubiquitous and Enduring in ML (implicit at times):

Traditional ML: Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

Modern ML: Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:

Which metric? Which values of k? Tradeoffs with big data?

A lot of recent insights towards these questions ...

k-Nearest Neighbor Approach:

Use the k closest datapoints to x *to infer* something about x.

Ubiquitous and Enduring in ML (implicit at times):

Traditional ML: Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

Modern ML: Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:

Which metric? Which values of k? Tradeoffs with big data?

A lot of recent insights towards these questions ...

k-Nearest Neighbor Approach:

Use the k closest datapoints to x *to infer* something about x.

Ubiquitous and Enduring in ML (implicit at times):

Traditional ML: Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

Modern ML: Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:

Which metric? Which values of k? Tradeoffs with big data?

A lot of recent insights towards these questions ...

k-Nearest Neighbor Approach:

Use the k closest datapoints to x *to infer* something about x.

Ubiquitous and Enduring in ML (implicit at times):

Traditional ML: Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

Modern ML: Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:

Which metric? Which values of k? Tradeoffs with big data?

A lot of recent insights towards these questions ...

k-Nearest Neighbor Approach:

Use the k closest datapoints to x *to infer* something about x.

Ubiquitous and Enduring in ML (implicit at times):

Traditional ML: Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

Modern ML: Matrix Completion, Inference on Graphs, Time Series Prediction ...

Of Practical Interest:

Which metric? Which values of k? Tradeoffs with big data?

A lot of recent insights towards these questions ...

k-Nearest Neighbor Approach:

Use the k closest datapoints to x *to infer* something about x.

Ubiquitous and Enduring in ML (implicit at times):

Traditional ML: Classification, Regression, Density Estimation, Bandits, Manifold Learning, Clustering ...

Modern ML: Matrix Completion, Inference on Graphs, Time Series Prediction ...

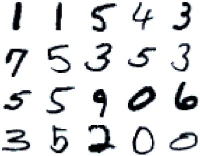
Of Practical Interest:

Which metric? Which values of k? Tradeoffs with big data?

A lot of recent insights towards these questions ...

Basic Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...



$\ldots \quad y \leftarrow 5$

Prediction: aggregate Y values in **Neighborhood**(x)

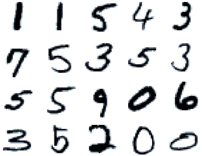
Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, Kpotufe, von Luxburg, Hein, Chaudhuri, Dasgupta, Langford, Kakade, Beygelzimer, Lee, Gray, Andoni, Clarkson, Krauthgamer, Indyk, ...

Basic Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...



$\ldots$ $x \equiv$  $\ldots$ $y \leftarrow 5$

Prediction: **aggregate** Y values in **Neighborhood**(x)

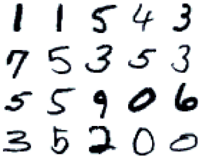
Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, Kpotufe, von Luxburg, Hein, Chaudhuri, Dasgupta, Langford, Kakade, Beygelzimer, Lee, Gray, Andoni, Clarkson, Krauthgamer, Indyk, ...

Basic Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...



$\ldots$ $x \equiv$  $\ldots$ $y \leftarrow 5$

Prediction: **aggregate** Y values in **Neighborhood**(x)

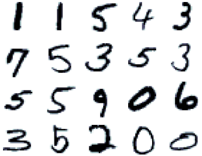
Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, Kpotufe, von Luxburg, Hein, Chaudhuri, Dasgupta, Langford, Kakade, Beygelzimer, Lee, Gray, Andoni, Clarkson, Krauthgamer, Indyk, ...

Basic Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...



$\ldots$ $x \equiv$

$\ldots$ $y \leftarrow 5$

Prediction: **aggregate** Y values in **Neighborhood**(x)

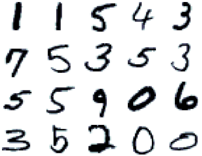
Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, Kpotufe, von Luxburg, Hein, Chaudhuri, Dasgupta, Langford, Kakade, Beygelzimer, Lee, Gray, Andoni, Clarkson, Krauthgamer, Indyk, ...

Basic Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...



$\ldots$ $x \equiv$  $\ldots$ $y \leftarrow 5$

Prediction: **aggregate** Y values in **Neighborhood**(x)

Similar Intuition: Classification Trees, RBF networks, Kernel machines.

Results by various authors help formalize the above intuition

Posner, Fix, Hodges, Cover, Hart, Devroye, Lugosi, Hero, Nobel, Györfi, Kulkarni, Ben David, Shalev-Schwartz, Samworth, Gadat, H. Chen, Shah, Kpotufe, von Luxburg, Hein, Chaudhuri, Dasgupta, Langford, Kakade, Beygelzimer, Lee, Gray, Andoni, Clarkson, Krauthgamer, Indyk, ...

Cover both Statistical and Algorithmic Issues:

1 **Statistical issues:** how well can NN perform?

- When is 1-NN enough?
- For k-NN, what should k be?
- Is there always a curse of dimension?

2 **Algorithmic issues:** how efficient can NN be?

- Which data structure to use?
- Can we parallelize NN?
- What do we tradeoff?

Cover both Statistical and Algorithmic Issues:

1 Statistical issues: how well can NN perform?

- When is 1-NN enough?
- For k-NN, what should k be?
- Is there always a curse of dimension?

2 Algorithmic issues: how efficient can NN be?

- Which data structure to use?
- Can we parallelize NN?
- What do we tradeoff?

Cover both Statistical and Algorithmic Issues:

1 **Statistical issues:** how well can NN perform?

- When is 1-NN enough?
- For k-NN, what should k be?
- Is there always a curse of dimension?

2 **Algorithmic issues:** how efficient can NN be?

- Which data structure to use?
- Can we parallelize NN?
- What do we tradeoff?

Data representation is important:

Examples:

- Direct Euclidean
- Deep Neural Representation (image, speech)
- Word Embedding (text)

$\cdots$

Representation $\equiv$ choice of metric or dissimilarity $\rho(x, x')$

Properties of ρ influence Statistical and Algorithmic aspects

...

Data representation is important:

Examples:

- Direct Euclidean
- Deep Neural Representation (image, speech)
- Word Embedding (text)

. . .

Representation ≡ choice of metric or dissimilarity $\rho(x, x')$

Properties of ρ influence Statistical and Algorithmic aspects

...

Data representation is important:

Examples:

- Direct Euclidean
- Deep Neural Representation (image, speech)
- Word Embedding (text)

$\ldots$

Representation $\equiv$ choice of metric or dissimilarity $\rho(x, x')$

Properties of ρ influence Statistical and Algorithmic aspects

...

Data representation is important:

Examples:

- Direct Euclidean
- Deep Neural Representation (image, speech)
- Word Embedding (text)

$\cdots$

Representation $\equiv$ choice of metric or dissimilarity $\rho(x, x')$

Properties of ρ influence Statistical and Algorithmic aspects

...

Tutorial Outline:

- **PART I:** Basic Statistical Insights
- **PART II:** Refined Analysis and Implementation

Tutorial Outline:

- **PART I:** Basic Statistical Insights
- **PART II:** Refined Analysis and Implementation

PART I: Basic Statistical Insights

- **Universality**
- Behavior of k-NN Distances
- From Regression to Classification
- Classification is easier than regression
- Multiclass and Mixed Costs

k-NN as a universal approach:

it can fit anything, provided k grows (but not too fast) with sample size!

Let's make this precise in the context of regression ...

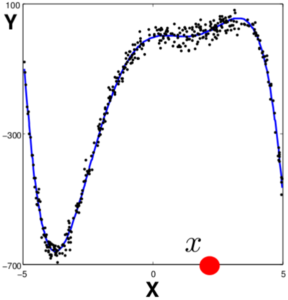
k-NN as a universal approach:

it can fit anything, provided k grows (but not too fast) with sample size!

Let's make this precise in the context of regression ...

k-NN Regression

i.i.d. Data: $\{(X_i, Y_i)\}_{i=1}^n$,
$Y = f(X) +$ noise



Learn: $f_k(x) =$ avg (Y_i) of k-NN(x).

k-NN is universally consistent:

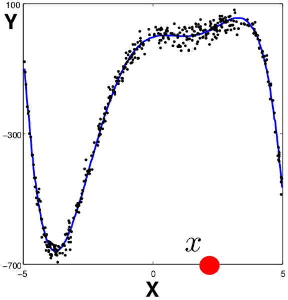
Suppose $\frac{k}{n} \to 0$ and $k \to \infty$, then $\mathbb{E}\,|f_k(X) - f(X)| \xrightarrow{n\to\infty} 0$

Any function f, no matter how complex.

k-NN Regression

i.i.d. Data: $\{(X_i, Y_i)\}_{i=1}^n$,
$Y = f(X) +$ noise

Learn: $f_k(x) =$ avg (Y_i) of k-NN(x).



k-NN is universally consistent:

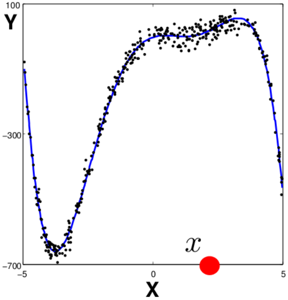
Suppose $\frac{k}{n} \to 0$ and $k \to \infty$, then $\mathbb{E}\,|f_k(X) - f(X)| \xrightarrow{n\to\infty} 0$

Any function f, no matter how complex.

k-NN Regression

i.i.d. Data: $\{(X_i, Y_i)\}_{i=1}^n$,
$Y = f(X) +$ noise

Learn: $f_k(x) =$ avg (Y_i) of k-NN(x).



k-NN is universally consistent:

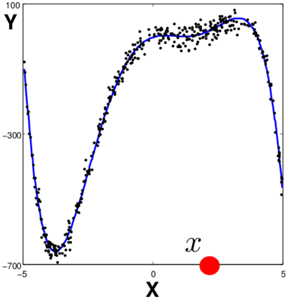
Suppose $\frac{k}{n} \to 0$ and $k \to \infty$, then $\mathbb{E}\,|f_k(X) - f(X)| \xrightarrow{n\to\infty} 0$

Any function f, no matter how complex.

k-NN Regression

i.i.d. Data: $\{(X_i, Y_i)\}_{i=1}^n$,
$Y = f(X) +$ noise

Learn: $f_k(x) =$ avg (Y_i) of k-NN(x).



k-NN is universally consistent:

Suppose $\frac{k}{n} \to 0$ and $k \to \infty$, then $\mathbb{E}\,|f_k(X) - f(X)| \xrightarrow{n\to\infty} 0$

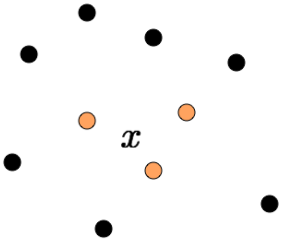
Any function f, no matter how complex.

Intuition:

- $\{X_{(i)}\}_1^k \to x$ as long as k is fixed or grows slow $(k/n \to 0)$
- Suppose f is continuous, then we also get $\{f(X_{(i)})\}_1^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k}\sum(f(X_{(i)} + \text{noise}) \to f(x)$

Now, any f can be approximated arbitrarily well by continuous f's $\square$

Intuition:

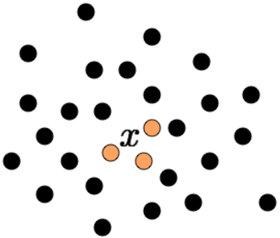


Consider the k-NN $\{X_{(i)}\}_1^k$ of some x

- $\{X_{(i)}\}_1^k \to x$ as long as k is fixed or grows slow ($k/n \to 0$)
- Suppose f is continuous, then we also get $\{f(X_{(i)})\}_1^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k}\sum(f(X_{(i)} + \text{noise}) \to f(x)$

Now, any f can be approximated arbitrarily well by continuous f's $\square$

Intuition:

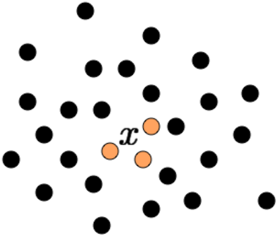


As $n \nearrow$, all $\{X_{(i)}\}_1^k$ move closer to x

- $\{X_{(i)}\}_1^k \to x$ as long as k is fixed or grows slow $(k/n \to 0)$
- Suppose f is continuous, then we also get $\{f(X_{(i)})\}_1^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k}\sum(f(X_{(i)} + \text{noise}) \to f(x)$

Now, any f can be approximated arbitrarily well by continuous f's $\square$

Intuition:

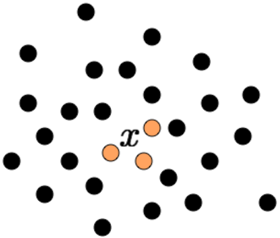


As $n \nearrow$, all $\{X_{(i)}\}_1^k$ move closer to x

- $\{X_{(i)}\}_1^k \to x$ as long as k is fixed or grows slow ($k/n \to 0$)
- Suppose f is continuous, then we also get $\{f(X_{(i)})\}_1^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k}\sum(f(X_{(i)}) + \text{noise}) \to f(x)$

Now, any f can be approximated arbitrarily well by continuous f's $\square$

Intuition:

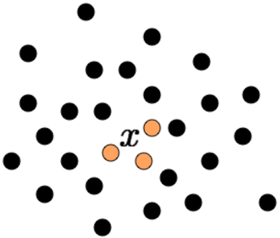


As $n \nearrow$, all $\{X_{(i)}\}_1^k$ move closer to x

- $\{X_{(i)}\}_1^k \to x$ as long as k is fixed or grows slow ($k/n \to 0$)
- Suppose f is continuous, then we also get $\{f(X_{(i)})\}_1^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k}\sum(f(X_{(i)} + \text{noise}) \to f(x)$

Now, any f can be approximated arbitrarily well by continuous f's $\square$

Intuition:

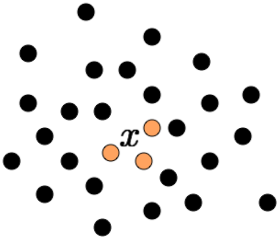


As $n \nearrow$, all $\{X_{(i)}\}_1^k$ move closer to x

- $\{X_{(i)}\}_1^k \to x$ as long as k is fixed or grows slow ($k/n \to 0$)
- Suppose f is continuous, then we also get $\{f(X_{(i)})\}_1^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k}\sum(f(X_{(i)} + \text{noise}) \to f(x)$

Now, any f can be approximated arbitrarily well by continuous f's $\square$

Intuition:



As $n \nearrow$, all $\{X_{(i)}\}_1^k$ move closer to x

- $\{X_{(i)}\}_1^k \to x$ as long as k is fixed or grows slow ($k/n \to 0$)
- Suppose f is continuous, then we also get $\{f(X_{(i)})\}_1^k \to f(x)$
- If $k \to \infty$, then $f_k(x) = \frac{1}{k}\sum(f(X_{(i)} + \text{noise}) \to f(x)$

Now, any f can be approximated arbitrarily well by continuous f's $\square$

Similar universality results for classification, density estimation, ...

Seminal results on k-NN consistency:

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity , $\mathbb{R}^d$.
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.

Main message: *k should grow (not too fast) with n ... (e.g. $k \sim \log n$)*

But we need a more refined picture ...

Similar universality results for classification, density estimation, ...

Seminal results on k-NN consistency:

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity , $\mathbb{R}^d$.
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.

Main message: *k should grow (not too fast) with n ... (e.g. $k \sim \log n$)*

But we need a more refined picture ...

Similar universality results for classification, density estimation, ...

Seminal results on k-NN consistency:

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity , $\mathbb{R}^d$.
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.

Main message: *k should grow (not too fast) with n ...* *(e.g. $k \sim \log n$)*

But we need a more refined picture ...

Similar universality results for classification, density estimation, ...

Seminal results on k-NN consistency:

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity , $\mathbb{R}^d$.
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.

Main message: *k should grow (not too fast) with n ... (e.g. $k \sim \log n$)*

But we need a more refined picture ...

Similar universality results for classification, density estimation, ...

Seminal results on k-NN consistency:

- [Fix, Hodges, 51]: classification + regularity, $\mathbb{R}^d$.
- [Cover, Hart, 65, 67, 68]: classification + regularity, any metric.
- [Stone, 77]: classification, universal, $\mathbb{R}^d$.
- [Devroye, Wagner, 77]: density estimation + regularity , $\mathbb{R}^d$.
- [Devroye, Gyorfi, Kryzak, Lugosi, 94]: regression, universal, $\mathbb{R}^d$.

Main message: *k should grow (not too fast) with n ... (e.g. $k \sim \log n$)*

But we need a more refined picture ...

PART I: Basic Statistical Insights

- Universality
- **Behavior of k-NN Distances**
- From Regression to Classification
- Classification is easier than regression
- Multiclass and Mixed Costs

Why k-NN Distances?

Recall Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...

So we hope that k-NN(x) are close to x ...

Formally: let $r_k(x) \equiv$ distance from x to k-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of (P_X, k, n)?

Why k-NN Distances?

Recall Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...

So we hope that k-NN(x) are close to x ...

Formally: let $r_k(x) \equiv$ distance from x to k-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of (P_X, k, n)?

Why k-NN Distances?

Recall Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...

So we hope that k-NN(x) are close to x ...

Formally: let $r_k(x) \equiv$ distance from x to k-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of (P_X, k, n)?

Why k-NN Distances?

Recall Intuition:

Closest neighbors of x should be mostly of similar type $y = y(x)$ ...

So we hope that k-NN(x) are close to x ...

Formally: let $r_k(x) \equiv$ distance from x to k-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of (P_X, k, n)?

Why k-NN Distances?

Recall Intuition:

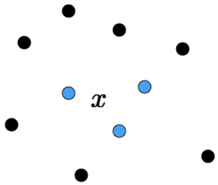
Closest neighbors of x should be mostly of similar type $y = y(x)$ ...

So we hope that k-NN(x) are close to x ...

Formally: let $r_k(x) \equiv$ distance from x to k-th NN in i.i.d. $\{X_i\}_1^n$

Q: How small is $r_k(x) \equiv$ function of (P_X, k, n)?

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



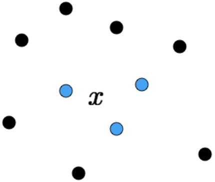
$B_x \equiv B(x, r_k(x)) \equiv$ smallest ball containing k-NN(x)

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$.

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



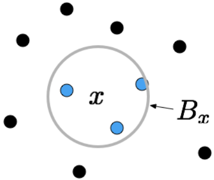
$B_x \equiv B(x, r_k(x)) \equiv$ smallest ball containing k-NN(x)

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$.

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



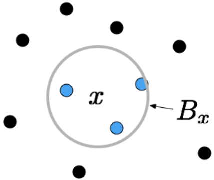
$B_x \equiv B(x, r_k(x)) \equiv$ smallest ball containing k-NN(x)

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

$$\text{Therefore, w.h.p.,} \quad r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}.$$

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



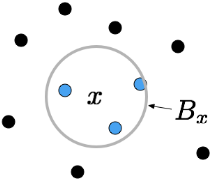
$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $\quad r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}.$

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



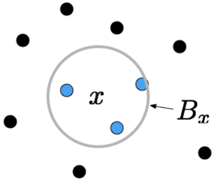
$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $\quad r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$.

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



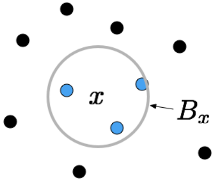
$B_x \equiv B(x, r_k(x)) \equiv$ smallest ball containing k-NN(x)

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$.

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



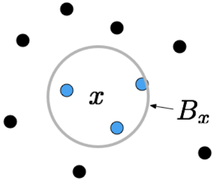
$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $\quad P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $\quad r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}$.

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



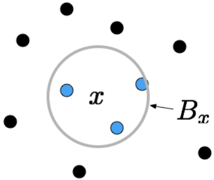
$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x')\, dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$.

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



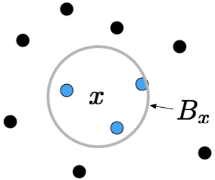
$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $\quad P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $\quad r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$.

Fix x, and assume $\{X_i\}_1^n$ i.i.d. P_X with density p_X in $\mathbb{R}^d$.



$$B_x \equiv B(x, r_k(x)) \equiv \text{smallest ball containing } k\text{-NN}(x)$$

- Assume no ties: $P_n(B_x) = k/n$.
- **w.h.p.** $P_n \approx P_X \implies P_X(B_x) \approx k/n$.

Now: $P_X(B_x) \equiv \int_{B_x} p_X(x')\,dx' \approx p_X(x) \cdot \int_{B_x} dx' \approx p_X(x) \cdot r_k(x)^d$.

Therefore, w.h.p., $\quad r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$.

$$r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$$

Immediate messages:

- $r_k(x) \nearrow$ when local density $p_X(x) \searrow$
- $r_k(x) \nearrow$ when input dimension $d \nearrow$

Use smaller k for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx (1/\epsilon)^d$ ...
Fortunately, effective d can be small for high-dimensional $X \in \mathbb{R}^D$

Effective d is whichever satisfies: $P_X(B(x, r)) = \ldots \approx c \cdot r^d$

$$d = d\left(\text{metric } \rho; \text{ where } X \text{ lies in } \mathbb{R}^D\right)$$

$$r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$$

Immediate messages:

- $r_k(x) \nearrow$ when local density $p_X(x) \searrow$
- $r_k(x) \nearrow$ when input dimension $d \nearrow$

Use smaller k for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx (1/\epsilon)^d$ ...

Fortunately, effective d can be small for high-dimensional $X \in \mathbb{R}^D$

Effective d is whichever satisfies: $P_X(B(x,r)) = \ldots \approx c \cdot r^d$

$$d = d\,(\text{metric } \rho; \text{ where } X \text{ lies in } \mathbb{R}^D)$$

$$r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$$

Immediate messages:

- $r_k(x) \nearrow$ when local density $p_X(x) \searrow$
- $r_k(x) \nearrow$ when input dimension $d \nearrow$

Use smaller k for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx (1/\epsilon)^d$ ...

Fortunately, effective d can be small for high-dimensional $X \in \mathbb{R}^D$

Effective d is whichever satisfies: $P_X(B(x,r)) = \ldots \approx c \cdot r^d$

$$d = d\,(\text{metric } \rho;\text{ where } X \text{ lies in } \mathbb{R}^D)$$

$$r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n} \right)^{1/d}$$

Immediate messages:

- $r_k(x) \nearrow$ when local density $p_X(x) \searrow$
- $r_k(x) \nearrow$ when input dimension $d \nearrow$

Use smaller k for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx (1/\epsilon)^d$ ...

Fortunately, effective d can be small for high-dimensional $X \in \mathbb{R}^D$

Effective d is whichever satisfies: $P_X(B(x, r)) = \ldots \approx c \cdot r^d$

$$d = d\left(\text{metric } \rho\text{; where } X \text{ lies in } \mathbb{R}^D\right)$$

$$r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$$

Immediate messages:

- $r_k(x) \nearrow$ when local density $p_X(x) \searrow$
- $r_k(x) \nearrow$ when input dimension $d \nearrow$

  Use smaller k for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx (1/\epsilon)^d$ ...

Fortunately, effective d can be small for high-dimensional $X \in \mathbb{R}^D$

Effective d is whichever satisfies: $P_X(B(x,r)) = \ldots \approx c \cdot r^d$

$$d = d\left(\text{metric } \rho;\text{ where } X \text{ lies in } \mathbb{R}^D\right)$$

$$r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$$

Immediate messages:

- $r_k(x) \nearrow$ when local density $p_X(x) \searrow$
- $r_k(x) \nearrow$ when input dimension $d \nearrow$

  Use smaller k for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx (1/\epsilon)^d$ ...

Fortunately, effective d can be small for high-dimensional $X \in \mathbb{R}^D$

Effective d is whichever satisfies: $P_X(B(x,r)) = \ldots \approx c \cdot r^d$

$$d = d\left(\text{metric } \rho\text{; where } X \text{ lies in } \mathbb{R}^D\right)$$

$$r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$$

Immediate messages:

- $r_k(x) \nearrow$ when local density $p_X(x) \searrow$
- $r_k(x) \nearrow$ when input dimension $d \nearrow$

Use smaller k for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx (1/\epsilon)^d$ ...

Fortunately, effective d can be small for high-dimensional $X \in \mathbb{R}^D$

Effective d is whichever satisfies: $P_X(B(x, r)) = \ldots \approx c \cdot r^d$

$d = d$ (metric ρ; where X lies in $\mathbb{R}^D$)

$$r_k(x) \approx \left(\frac{1}{p_X(x)} \cdot \frac{k}{n}\right)^{1/d}$$

Immediate messages:

- $r_k(x) \nearrow$ when local density $p_X(x) \searrow$
- $r_k(x) \nearrow$ when input dimension $d \nearrow$

  Use smaller k for higher dimensional data ...

Curse of dimension: For $r_k \approx \epsilon$ we need $n \approx (1/\epsilon)^d$ ...

Fortunately, effective d can be small for high-dimensional $X \in \mathbb{R}^D$

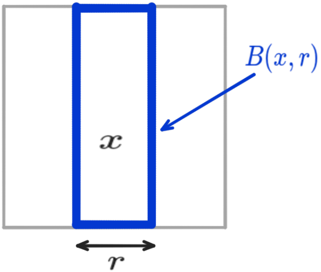
Effective d is whichever satisfies: $P_X(B(x, r)) = \ldots \approx c \cdot r^d$

$$d = d\,(\text{metric } \rho\text{; where } X \text{ lies in } \mathbb{R}^D)$$

Ex 1: Suppose $X \in \mathbb{R}^D$, but $\rho(x, x') \approx \rho(x_{(1)}, x'_{(1)})$ ...

$$P_X(B(x,r)) \approx r \implies \text{effective } d = 1$$

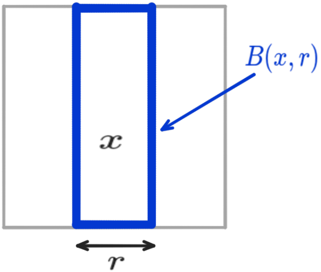
Ex 1: Suppose $X \in \mathbb{R}^D$, but $\rho(x, x') \approx \rho(x_{(1)}, x'_{(1)})$ ...



$$B(x, r) \equiv \{x' : \rho(x, x') \leq r\}$$

$P_X(B(x, r)) \approx r \implies$ effective $d = 1$

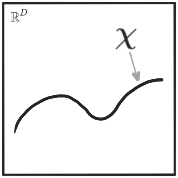
Ex 1: Suppose $X \in \mathbb{R}^D$, but $\rho(x, x') \approx \rho(x_{(1)}, x'_{(1)})$ ...



$$B(x, r) \equiv \{x' : \rho(x, x') \leq r\}$$

$$P_X(B(x, r)) \approx r \implies \text{effective } d = 1$$

Ex 2: Suppose $X \in \mathbb{R}^D$, but lies on a d-dimensional space $\mathcal{X}$ ...

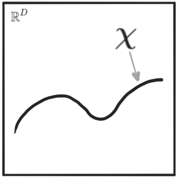


Consider B, of radius r, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

Thus we'd have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$.

Ex 2: Suppose $X \in \mathbb{R}^D$, but lies on a d-dimensional space $\mathcal{X}$ ...

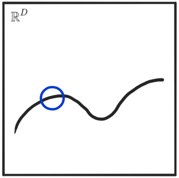


Consider B, of radius r, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

Thus we'd have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$.

Ex 2: Suppose $X \in \mathbb{R}^D$, but lies on a d-dimensional space $\mathcal{X}$ ...

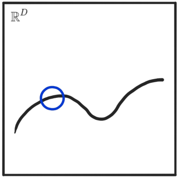


Consider B, of radius r, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

Thus we'd have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$.

Ex 2: Suppose $X \in \mathbb{R}^D$, but lies on a d-dimensional space $\mathcal{X}$ ...

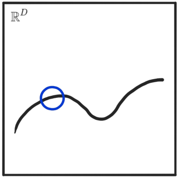


Consider B, of radius r, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

Thus we'd have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$.

Ex 2: Suppose $X \in \mathbb{R}^D$, but lies on a d-dimensional space $\mathcal{X}$ ...

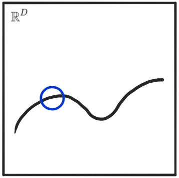


Consider B, of radius r, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

Thus we'd have $r_k(x) \approx (k/n)^{1/d}$, irrespective of $D \gg d$.

Ex 2: Suppose $X \in \mathbb{R}^D$, but lies on a d-dimensional space $\mathcal{X}$ ...

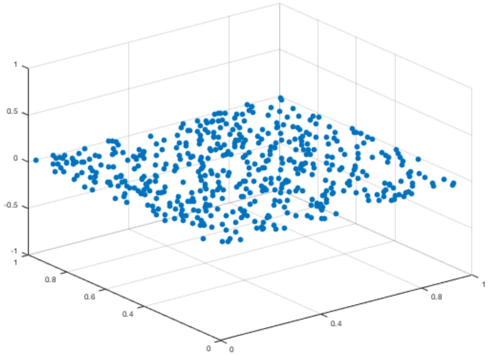


Consider B, of radius r, centered on $\mathcal{X}$:

$$P_X(B) \approx p_X \cdot \int_{B \cap \mathcal{X}} dx \approx p_X \cdot r^d$$

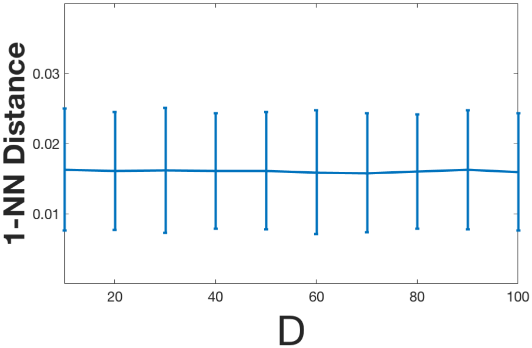
Thus we'd have $r_k(x) \approx (k/n)^{1/d}$**, irrespective of** $D \gg d$**.**

Quick Simulations:



Embed $(d = 2)$-data into high-dimensional $\mathbb{R}^D$, $D \to \infty$

Quick Simulations:



Fix $d = 2$: average NN distances are stable as D varies

Refined analysis for $r_k(x)$:

[J. Costa, A. Hero 04], [R. Samworth 12]

Implications:

$r_k(x)$ **adaptive to** d $\implies$ **NN methods adaptive to** d **...**

(d**-sparse documents, images, Robotics data on** d**-manifold)**

Refined analysis for $r_k(x)$:

[J. Costa, A. Hero 04], [R. Samworth 12]

Implications:

$r_k(x)$ **adaptive to** d $\implies$ **NN methods adaptive to** d **...**

(d**-sparse documents, images, Robotics data on** d**-manifold)**

PART I: Basic Statistical Insights

- Universality
- Behavior of k-NN Distances
- **From Regression to Classification**
- Classification is easier than regression
- Multiclass and Mixed Costs

From bounds on $r_k(x)$ to error rates:

Program:

1. Regression bounds
2. Reduce Classification to Regression

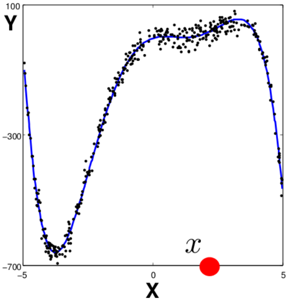
From bounds on $r_k(x)$ to error rates:

Program:

1. Regression bounds
2. Reduce Classification to Regression

k-NN Regression

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y = f(X) + \text{noise}$



Learn: $f_k(x) = \text{avg } (Y_i)$ of k-NN(x).

Ideal Metric ρ**:** $f(x) \approx f(x')$ if $\rho(x, x') \approx 0$

... e.g., assume f is Lipschitz: $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')$.

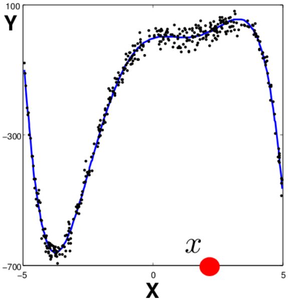
Performance Goal:

Pick k such that $\|f_k - f\|^2 \equiv \mathbb{E}_X |f_k(X) - f(X)|^2$ is small.

k-NN Regression

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y = f(X) + \text{noise}$

Learn: $f_k(x) = \text{avg } (Y_i) \text{ of } k\text{-NN}(x)$.



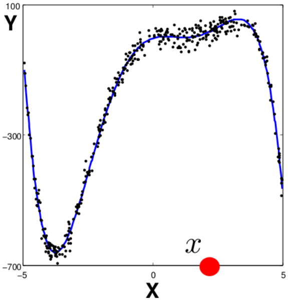
Ideal Metric ρ**:** $f(x) \approx f(x')$ if $\rho(x, x') \approx 0$

... e.g., assume f is Lipschitz: $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')$.

Performance Goal:

Pick k such that $\|f_k - f\|^2 \equiv \mathbb{E}_X |f_k(X) - f(X)|^2$ is small.

k-NN Regression



Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y = f(X) + \text{noise}$

Learn: $f_k(x) = \text{avg } (Y_i) \text{ of } k\text{-NN}(x)$.

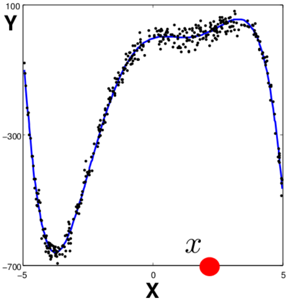
Ideal Metric ρ: $f(x) \approx f(x')$ if $\rho(x, x') \approx 0$

... e.g., assume f is Lipschitz: $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')$.

Performance Goal:

Pick k such that $\|f_k - f\|^2 \equiv \mathbb{E}_X |f_k(X) - f(X)|^2$ is small.

k-NN Regression



Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y = f(X) + \text{noise}$

Learn: $f_k(x) = \text{avg } (Y_i) \text{ of } k\text{-NN}(x)$.

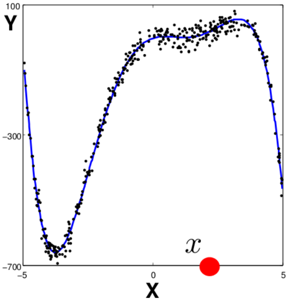
Ideal Metric ρ: $f(x) \approx f(x')$ if $\rho(x, x') \approx 0$

... e.g., assume f is Lipschitz: $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')$.

Performance Goal:

Pick k such that $\|f_k - f\|^2 \equiv \mathbb{E}_X |f_k(X) - f(X)|^2$ is small.

k-NN Regression



Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y = f(X) + \text{noise}$

Learn: $f_k(x) = \text{avg } (Y_i)$ of k-NN(x).

Ideal Metric ρ: $f(x) \approx f(x')$ if $\rho(x, x') \approx 0$

... e.g., assume f is Lipschitz: $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')$.

Performance Goal:

Pick k such that $\|f_k - f\|^2 \equiv \mathbb{E}_X \, |f_k(X) - f(X)|^2$ is small.

Step 1: *Bias-variance decomposition*

A simple fact: $\mathbb{E}\,|Z-c|^2 = \mathbb{E}\,|Z-\mathbb{E}Z|^2 + |c-\mathbb{E}Z|^2$.

So fix x, and fix $\{X_i\}$, and let $\tilde{f}_k(x) = \mathbb{E}_{\{Y_i\}} f_k(x)$ ...

$$\mathbb{E}\,|f_k(x)-f(x)|^2 = \underbrace{\mathbb{E}\,|f_k(x)-\tilde{f}_k(x)|^2}_{\text{Variance}} + \underbrace{|f(x)-\tilde{f}_k(x)|^2}_{\text{Bias}^2}.$$

Step 1: *Bias-variance decomposition*

A simple fact: $\mathbb{E}\left|Z - c\right|^2 = \mathbb{E}\left|Z - \mathbb{E}Z\right|^2 + \left|c - \mathbb{E}Z\right|^2.$

So fix x, and fix $\{X_i\}$, and let $\tilde{f}_k(x) = \mathbb{E}_{\{Y_i\}} f_k(x)$ ...

$$\mathbb{E}\left|f_k(x) - f(x)\right|^2 = \underbrace{\mathbb{E}\left|f_k(x) - \tilde{f}_k(x)\right|^2}_{\text{Variance}} + \underbrace{\left|f(x) - \tilde{f}_k(x)\right|^2}_{\text{Bias}^2}.$$

Step 1: *Bias-variance decomposition*

A simple fact: $\mathbb{E}\left|Z-c\right|^2 = \mathbb{E}\left|Z-\mathbb{E}Z\right|^2 + \left|c-\mathbb{E}Z\right|^2.$

So fix x, and fix $\{X_i\}$, and let $\tilde{f}_k(x) = \mathbb{E}_{\{Y_i\}} f_k(x)$ ...

$$\mathbb{E}\left|f_k(x) - f(x)\right|^2 = \underbrace{\mathbb{E}\left|f_k(x) - \tilde{f}_k(x)\right|^2}_{\text{Variance}} + \underbrace{\left|f(x) - \tilde{f}_k(x)\right|^2}_{\text{Bias}^2}.$$

Step 1: *Bias-variance decomposition*

A simple fact: $\mathbb{E}\left|Z-c\right|^2 = \mathbb{E}\left|Z-\mathbb{E}Z\right|^2 + \left|c-\mathbb{E}Z\right|^2.$

So fix x, and fix $\{X_i\}$, and let $\tilde{f}_k(x) = \mathbb{E}_{\{Y_i\}} f_k(x)$ ...

$$\mathbb{E}\left|f_k(x)-f(x)\right|^2 = \underbrace{\mathbb{E}\left|f_k(x)-\tilde{f}_k(x)\right|^2}_{\text{Variance}} + \underbrace{\left|f(x)-\tilde{f}_k(x)\right|^2}_{\text{Bias}^2}.$$

Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} Y_i$

$$\mathrm{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k\text{-NN}(x)} \mathrm{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} f(X_i)$

$$\begin{aligned}\left|\tilde{f}_k(x) - f(x)\right| &\le \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} |f(X_i) - f(x))| \\ &\le \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} \rho(X_i, x) \\ &\le r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}.\end{aligned}$$

Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} Y_i$

$$\mathrm{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k\text{-NN}(x)} \mathrm{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} f(X_i)$

$$\begin{aligned}\left|\tilde{f}_k(x) - f(x)\right| &\leq \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} |f(X_i) - f(x))| \\ &\leq \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} \rho(X_i, x) \\ &\leq r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}.\end{aligned}$$

Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} Y_i$

$$\mathrm{Var}(f_k(x)) = \frac{1}{k^2}\sum_{X_i \in k\text{-NN}(x)} \mathrm{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} f(X_i)$

$$\begin{aligned}
\left|\tilde{f}_k(x) - f(x)\right| &\le \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} |f(X_i) - f(x))| \\
&\le \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} \rho(X_i, x) \\
&\le r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}.
\end{aligned}$$

Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} Y_i$

$$\mathrm{Var}(f_k(x)) = \frac{1}{k^2}\sum_{X_i \in k\text{-NN}(x)} \mathrm{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} f(X_i)$

$$\begin{aligned}\left|\tilde{f}_k(x) - f(x)\right| &\le \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} |f(X_i) - f(x))| \\ &\le \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} \rho(X_i, x) \\ &\le r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}.\end{aligned}$$

Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} Y_i$

$$\mathrm{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k\text{-NN}(x)} \mathrm{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} f(X_i)$

$$\begin{aligned}
\left|\tilde{f}_k(x) - f(x)\right| &\le \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} |f(X_i) - f(x)| \\
&\le \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} \rho(X_i, x) \\
&\le r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}.
\end{aligned}$$

Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} Y_i$

$$\mathsf{Var}(f_k(x)) = \frac{1}{k^2} \sum_{X_i \in k\text{-NN}(x)} \mathsf{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} f(X_i)$

$$\begin{aligned}
\left|\tilde{f}_k(x) - f(x)\right| &\le \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} |f(X_i) - f(x))| \\
&\le \frac{1}{k} \sum_{X_i \in k\text{-NN}(x)} \rho(X_i, x) \\
&\le r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}.
\end{aligned}$$

Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} Y_i$

$$\mathrm{Var}(f_k(x)) = \frac{1}{k^2}\sum_{X_i \in k\text{-NN}(x)} \mathrm{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} f(X_i)$

$$\begin{aligned} \left|\tilde{f}_k(x) - f(x)\right| &\leq \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} |f(X_i) - f(x))| \\ &\leq \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} \rho(X_i, x) \\ &\leq r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}. \end{aligned}$$

Step 2: *Bound the two terms*

- **Variance:** recall $f_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} Y_i$

$$\text{Var}(f_k(x)) = \frac{1}{k^2}\sum_{X_i \in k\text{-NN}(x)} \text{Var}(Y_i) = \frac{\sigma_Y^2}{k}$$

- **Bias:** note that $\tilde{f}_k(x) = \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} f(X_i)$

$$\begin{aligned}\left|\tilde{f}_k(x) - f(x)\right| &\leq \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} |f(X_i) - f(x)| \\ &\leq \frac{1}{k}\sum_{X_i \in k\text{-NN}(x)} \rho(X_i, x) \\ &\leq r_k(x) \approx \left(\frac{k}{n}\right)^{1/d}.\end{aligned}$$

Step 3: *Integrate over x and $\{X_i\}$*

We then get: $\mathbb{E}\,\|f_k - f\|^2 \lesssim \dfrac{1}{k} + \left(\dfrac{k}{n}\right)^{2/d}.$

Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E}\,\|f_k - f\|^2 \lesssim n^{-2/(2+d)}$, optimal!

Best choice of $k \nearrow$ as $n \nearrow$ and $d \searrow$

Choosing k by C-V yields same optimal rates.

Step 3: *Integrate over x and $\{X_i\}$*

We then get: $$\mathbb{E}\,\|f_k - f\|^2 \lesssim \frac{1}{k} + \left(\frac{k}{n}\right)^{2/d}.$$

Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E}\,\|f_k - f\|^2 \lesssim n^{-2/(2+d)}$, optimal.

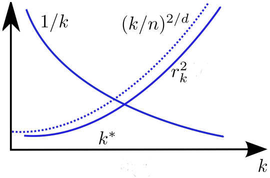
Best choice of $k \nearrow$ as $n \nearrow$ and $d \searrow$

Choosing k by C-V yields same optimal rates.

Step 3: *Integrate over x and $\{X_i\}$*

We then get: $\mathbb{E}\,\|f_k - f\|^2 \lesssim \frac{1}{k} + \left(\frac{k}{n}\right)^{2/d}.$

Tradeoff on k:



Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E}\,\|f_k - f\|^2 \lesssim n^{-2/(2+d)}$, optimal.

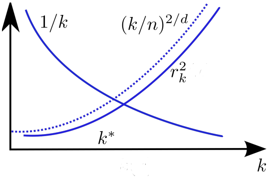
Best choice of $k \nearrow$ as $n \nearrow$ and $d \searrow$

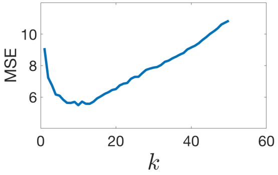
Choosing k by C-V yields same optimal rates.

Step 3: *Integrate over x and $\{X_i\}$*

We then get: $\mathbb{E}\,\|f_k - f\|^2 \lesssim \dfrac{1}{k} + \left(\dfrac{k}{n}\right)^{2/d}.$

Tradeoff on k:





Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E}\,\|f_k - f\|^2 \lesssim n^{-2/(2+d)}$, optimal.

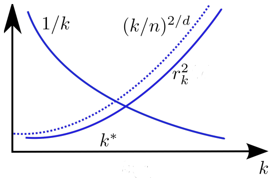
Best choice of $k \nearrow$ as $n \nearrow$ and $d \searrow$

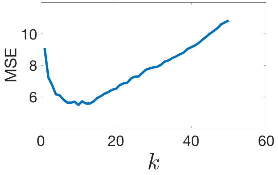
Choosing k by C-V yields same optimal rates.

Step 3: *Integrate over x and $\{X_i\}$*

We then get: $$\mathbb{E}\,\|f_k - f\|^2 \lesssim \frac{1}{k} + \left(\frac{k}{n}\right)^{2/d}.$$

Tradeoff on k:





Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E}\,\|f_k - f\|^2 \lesssim n^{-2/(2+d)}$, optimal.

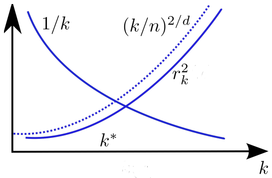
Best choice of $k \nearrow$ as $n \nearrow$ and $d \searrow$

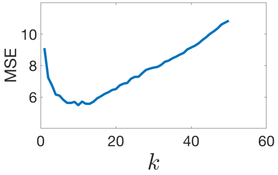
Choosing k by C-V yields same optimal rates.

Step 3: *Integrate over x and $\{X_i\}$*

We then get: $\mathbb{E}\,\|f_k - f\|^2 \lesssim \dfrac{1}{k} + \left(\dfrac{k}{n}\right)^{2/d}.$

Tradeoff on k:





Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E}\,\|f_k - f\|^2 \lesssim n^{-2/(2+d)}$, optimal.

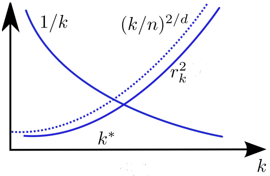
Best choice of $k \nearrow$ as $n \nearrow$ and $d \searrow$

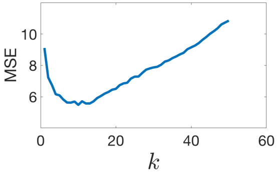
Choosing k by C-V yields same optimal rates.

Step 3: *Integrate over x and $\{X_i\}$*

We then get: $\mathbb{E}\,\|f_k - f\|^2 \lesssim \dfrac{1}{k} + \left(\dfrac{k}{n}\right)^{2/d}.$

Tradeoff on k:





Pick $k = \Theta(n^{2/(2+d)})$ to get $\mathbb{E}\,\|f_k - f\|^2 \lesssim n^{-2/(2+d)}$, optimal.

Best choice of $k \nearrow$ as $n \nearrow$ and $d \searrow$

Choosing k by C-V yields same optimal rates.

Unbounded capacity but generalizes at non-trivial rates ...

True even when k is different at every $x \in \mathbb{R}^d$
(infinite number of parameters) [Kpo. 11]

Unbounded capacity but generalizes at non-trivial rates ...

True even when k is different at every $x \in \mathbb{R}^d$

(infinite number of parameters) [Kpo. 11]

Similar messages under generalizations of Lipschitz assumption:

- **Hölder continuity:** $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^\alpha$.

(avg. version leads to so-called Nikoskii, Sobolev conditions)

Similar messages under generalizations of Lipschitz assumption:

- **Hölder continuity:** $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^{\alpha}$.

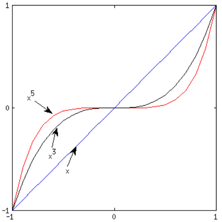
(avg. version leads to so-called Nikoskii, Sobolev conditions)

(see e.g. [Györfi, Krzyżak, Walk, 02])

Similar messages under generalizations of Lipschitz assumption:

- **Hölder continuity:** $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^{\alpha}$.

(avg. version leads to so-called Nikoskii, Sobolev conditions)



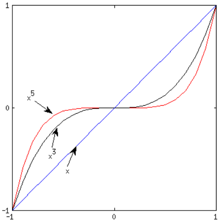
$|x - x'|^{\alpha}$ gets *flatter* around $x = 0$ as $\alpha \nearrow$.

(see e.g. [Györfi, Krzyżak, Walk, 02])

Similar messages under generalizations of Lipschitz assumption:

- **Hölder continuity:** $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^\alpha$.

(avg. version leads to so-called Nikoskii, Sobolev conditions)



$|x - x'|^\alpha$ gets *flatter* around $x = 0$ as $\alpha \nearrow$.

Additional messages (as $\alpha \nearrow$):

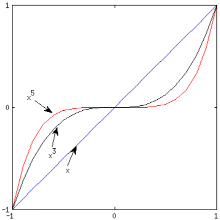
- Local averages (as k-NN) not appropriate for smoother (easier) f.
- Local polynomials are best, but harder to implement in high-D.

(see e.g. [Györfi, Krzyżak, Walk, 02])

Similar messages under generalizations of Lipschitz assumption:

- **Hölder continuity:** $|f(x) - f(x')| \leq \lambda \cdot \rho(x, x')^\alpha$.

(avg. version leads to so-called Nikoskii, Sobolev conditions)



$|x - x'|^\alpha$ gets *flatter* around $x = 0$ as $\alpha \nearrow$.

Additional messages (as $\alpha \nearrow$):

- Local averages (as k-NN) not appropriate for smoother (easier) f.
- Local polynomials are best, but harder to implement in high-D.

(see e.g. [Györfi, Krzyżak, Walk, 02])

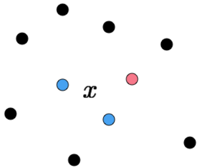
From bounds on $r_k(x)$ to error rates:

Program:

1. ~~Regression bounds~~
2. Reduce Classification to Regression

k-NN Classification

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{0, 1\}$.



Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).

Reduces to regression: let $f_k(x) =$ avg (Y_i) of k-NN(x)

... then: $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.

Performance Goal:

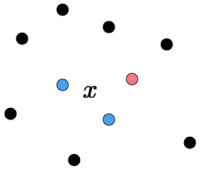
Pick k such that $\text{err}(h_k) \doteq \mathbb{P}(h_k(X) \neq Y)$ is small.

Equivalently, consider $\mathcal{E}(h_k) = \text{err}(h_k) - \text{err}(h^*)$.

k-NN Classification

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{0, 1\}$.

Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).



Reduces to regression: let $f_k(x) =$ avg (Y_i) of k-NN(x)

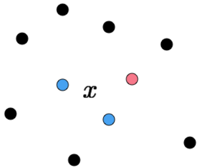
... then: $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.

Performance Goal:

Pick k such that $\text{err}(h_k) \equiv \mathbb{P}(h_k(X) \neq Y)$ is small.

Equivalently, consider $\mathcal{E}(h_k) = \text{err}(h_k) - \text{err}(h^*)$.

k-NN Classification



Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{0, 1\}$.

Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).

Reduces to regression: let $f_k(x) =$ avg (Y_i) of k-NN(x)

... then: $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.

Performance Goal:

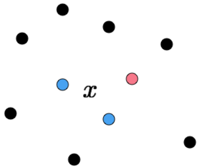
Pick k such that $\text{err}(h_k) \equiv \mathbb{P}(h_k(X) \neq Y)$ is small.

Equivalently, consider $\mathcal{E}(h_k) = \text{err}(h_k) - \text{err}(h^*)$.

k-NN Classification

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{0, 1\}$.

Learn: $h_k(x)$ = majority (Y_i) of k-NN(x).



Reduces to regression: let $f_k(x)$ = avg (Y_i) of k-NN(x)

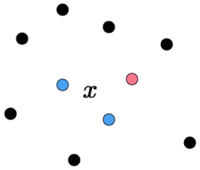
... then: $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.

Performance Goal:

Pick k such that err$(h_k) \equiv \mathbb{P}(h_k(X) \neq Y)$ is small.

Equivalently, consider $\mathcal{E}(h_k) =$ err$(h_k) -$ err(h^*).

k-NN Classification



Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{0, 1\}$.

Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).

Reduces to regression: let $f_k(x) =$ avg (Y_i) of k-NN(x)

... then: $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.

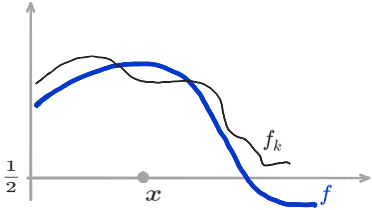
Performance Goal:

Pick k such that err$(h_k) \equiv \mathbb{P}(h_k(X) \neq Y)$ is small.

Equivalently, consider $\mathcal{E}(h_k) =$ err$(h_k) -$ err(h^*).

Remarks: $f_k(x)$ estimates $f(x) = \mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x)$,

and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.



$f_k(x) \approx f(x)$ implies $h_k(x) = h^*(x)$

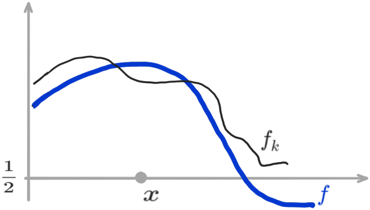
One can show: $\mathcal{E}(h_k) \leq 2\,\|f_k - f\|$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of k ...

Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}\,[Y|x] = \mathbb{P}(Y = 1|x)$,

and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.



$f_k(x) \approx f(x)$ implies $h_k(x) = h^*(x)$

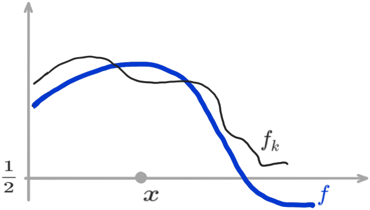
One can show: $\mathcal{E}(h_k) \leq 2\,\|f_k - f\|$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of k ...

Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}\left[Y|x\right] = \mathbb{P}(Y = 1|x)$,

and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.



$f_k(x) \approx f(x)$ implies $h_k(x) = h^*(x)$

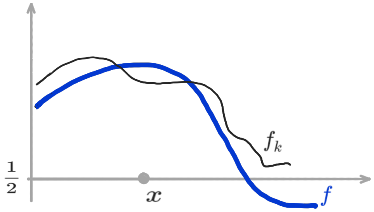
One can show: $\mathcal{E}(h_k) \leq 2\left\|f_k - f\right\|$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of k ...

Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x)$,

and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.



$f_k(x) \approx f(x)$ implies $h_k(x) = h^*(x)$

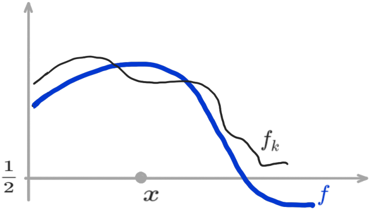
One can show: $\mathcal{E}(h_k) \leq 2\,\|f_k - f\|$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of k ...

Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x)$,

and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.



$f_k(x) \approx f(x)$ implies $h_k(x) = h^*(x)$

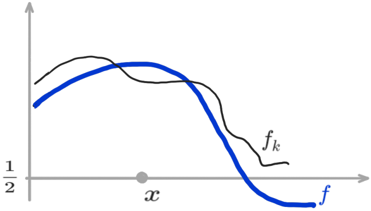
One can show: $\mathcal{E}(h_k) \leq 2\,\|f_k - f\|$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of k ...

Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}\,[Y|x] = \mathbb{P}(Y = 1|x)$,

and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.



$f_k(x) \approx f(x)$ implies $h_k(x) = h^*(x)$

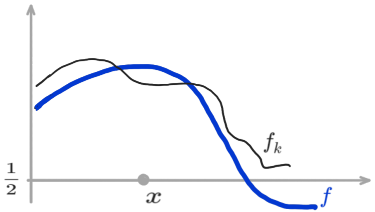
One can show: $\mathcal{E}(h_k) \leq 2\,\|f_k - f\|$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of k ...

Remarks: $f_k(x)$ estimates $f(x) \equiv \mathbb{E}[Y|x] = \mathbb{P}(Y = 1|x)$,

and $h^*(x) = \mathbb{1}\{f(x) \geq 1/2\}$, while $h_k(x) \equiv \mathbb{1}\{f_k(x) \geq 1/2\}$.



$f_k(x) \approx f(x)$ implies $h_k(x) = h^*(x)$

One can show: $\mathcal{E}(h_k) \leq 2 \|f_k - f\|$.

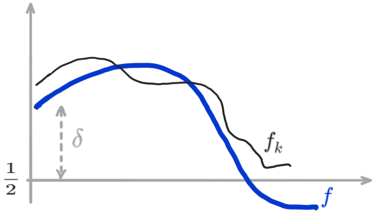
For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-1/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Similar messages on choice of k ...

PART I: Basic Statistical Insights

- Universality
- Behavior of k-NN Distances
- From Regression to Classification
- **Classification is easier than regression**
- Multiclass and Mixed Costs

$h^* = \mathbb{1}\{f \geq 1/2\}$, while $h_k \equiv \mathbb{1}\{f_k \geq 1/2\}$.



Suppose $|f(x) - 1/2| \geq \delta$ for most values x ...

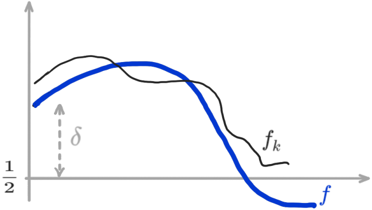
Then $|f_k - f| < \delta$ implies $h_k = h^*$ often ... no need for $f_k \approx f$.

Tsybakov's noise condition: $\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta$

If $|f_k - f| < \delta_n$, then $\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

$h^* = \mathbb{1}\{f \geq 1/2\}$, while $h_k \equiv \mathbb{1}\{f_k \geq 1/2\}$.



Suppose $|f(x) - 1/2| \geq \delta$ for most values x ...

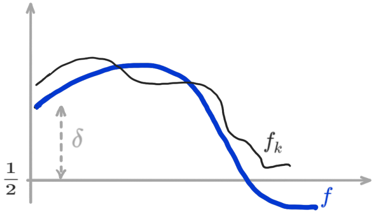
Then $|f_k - f| < \delta$ implies $h_k = h^*$ often ... no need for $f_k \approx f$.

Tsybakov's noise condition: $\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta$

If $|f_k - f| < \delta_n$, then $\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

$h^* = \mathbb{1}\{f \geq 1/2\}$, while $h_k \equiv \mathbb{1}\{f_k \geq 1/2\}$.



Suppose $|f(x) - 1/2| \geq \delta$ for most values x ...

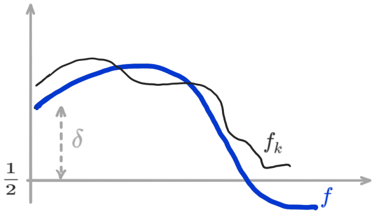
Then $|f_k - f| < \delta$ implies $h_k = h^*$ often ... <u>no need for $f_k \approx f$</u>.

Tsybakov's noise condition: $\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta$

If $|f_k - f| < \delta_n$, then $\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

$h^* = \mathbb{1}\{f \geq 1/2\}$, while $h_k \equiv \mathbb{1}\{f_k \geq 1/2\}$.



Suppose $|f(x) - 1/2| \geq \delta$ for most values x ...

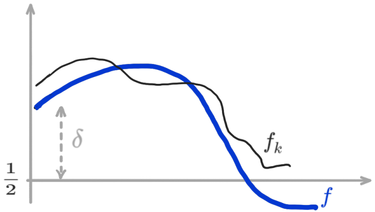
Then $|f_k - f| < \delta$ implies $h_k = h^*$ often ... no need for $f_k \approx f$.

Tsybakov's noise condition: $\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta$

If $|f_k - f| < \delta_n$, then $\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

$h^* = \mathbb{1}\{f \geq 1/2\}$, while $h_k \equiv \mathbb{1}\{f_k \geq 1/2\}$.



Suppose $|f(x) - 1/2| \geq \delta$ for most values x ...

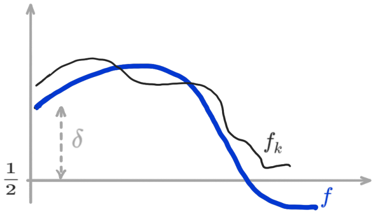
Then $|f_k - f| < \delta$ implies $h_k = h^*$ often ... no need for $f_k \approx f$.

Tsybakov's noise condition: $\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta$

If $|f_k - f| < \delta_n$, then $\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

$h^* = \mathbb{1}\{f \geq 1/2\}$, while $h_k \equiv \mathbb{1}\{f_k \geq 1/2\}$.



Suppose $|f(x) - 1/2| \geq \delta$ for most values x ...

Then $|f_k - f| < \delta$ implies $h_k = h^*$ often ... no need for $f_k \approx f$.

Tsybakov's noise condition: $\mathbb{P}_X(|f - 1/2| < \delta) \leq \delta^\beta$

If $|f_k - f| < \delta_n$, then $\mathbb{P}_X(h_k \neq h^*) \leq \mathbb{P}_X(|f - 1/2| < \delta_n) \leq \delta_n^\beta$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of f, and intrinsic d.
- Large margin β mitigates effects of metric.
  ($\beta = \infty \implies$ no curse of dimension!)

Technical Remarks:

- Above rates assume $P_X \equiv$ Uniform.
([Chaudhuri, Dasgupta 14] [Gadat et al 14]).

- For non-uniform P_X, rates are worse, but understudied.
([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of f, and intrinsic d.
- Large margin β mitigates effects of metric.
  ($\beta = \infty \implies$ no curse of dimension!)

Technical Remarks:

- Above rates assume $P_X \equiv$ Uniform.
([Chaudhuri, Dasgupta 14] [Gadat et al 14]).

- For non-uniform P_X, rates are worse, but understudied.
([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of f, and intrinsic d.
- Large margin β mitigates effects of metric.
  ($\beta = \infty \implies$ no curse of dimension!)

Technical Remarks:

- Above rates assume $P_X \equiv$ Uniform.
([Chaudhuri, Dasgupta 14] [Gadat et al 14]).
- For non-uniform P_X, rates are worse, but understudied.
([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of f, and intrinsic d.
- Large margin β mitigates effects of metric.
  ($\beta = \infty \implies$ no curse of dimension!)

Technical Remarks:

- Above rates assume $P_X \equiv$ Uniform.
([Chaudhuri, Dasgupta 14] [Gadat et al 14]).

- For non-uniform P_X, rates are worse, but understudied.
([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).

For Lipschitz f: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim n^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

- Choice of metric $\rightarrow$ Lipschitzness of f, and intrinsic d.
- Large margin β mitigates effects of metric.
  ($\beta = \infty \implies$ no curse of dimension!)

Technical Remarks:

- Above rates assume $P_X \equiv$ Uniform.

([Chaudhuri, Dasgupta 14] [Gadat et al 14]).

- For non-uniform P_X, rates are worse, but understudied.

([Gadat et al 14], [Cannings et al 17], [Kpo., Martinet 17]).

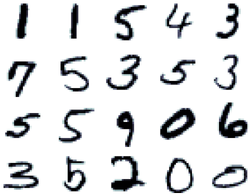
PART I: Basic Statistical Insights

- Universality
- Behavior of k-NN Distances
- From Regression to Classification
- Classification is easier than regression
- **Multiclass and Mixed Costs**

k-NN extends naturally to multiclass

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{1, , \ldots, L\}$.

Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).



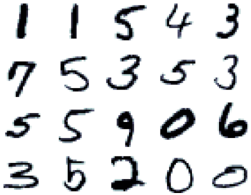
Reduction: let $f_k^y(x) =$ proportion $(Y = y)$ out of k-NN(x)

It estimates $f^y(x) = \mathbb{P}(Y = y|x)$.

... then: $h_k(x) \equiv \text{argmax}_y\{f_k^y(x)\}$, and $h^*(x) = \text{argmax}_y\{f^y(x)\}$

Previous insights extend easily ...

k-NN extends naturally to multiclass



Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{1, , \ldots, L\}$.

Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).

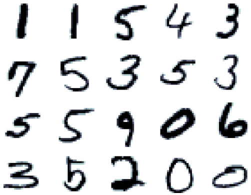
Reduction: let $f_k^y(x) =$ proportion $(Y = y)$ out of k-NN(x)

It estimates $f^y(x) = \mathbb{P}(Y = y|x)$.

... then: $h_k(x) \equiv \mathrm{argmax}_y\{f_k^y(x)\}$, and $h^*(x) = \mathrm{argmax}_y\{f^y(x)\}$

Previous insights extend easily ...

k-NN extends naturally to multiclass



Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{1, , \ldots, L\}$.

Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).

Reduction: let $f_k^y(x) =$ proportion $(Y = y)$ out of k-NN(x)

It estimates $f^y(x) = \mathbb{P}(Y = y|x)$.

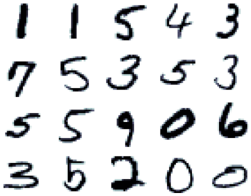
... then: $h_k(x) \equiv \text{argmax}_y\{f_k^y(x)\}$, and $h^*(x) = \text{argmax}_y\{f^y(x)\}$

Previous insights extend easily ...

k-NN extends naturally to multiclass

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{1, , \ldots, L\}$.

Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).



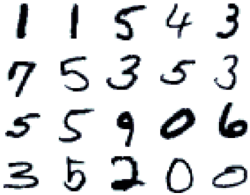
Reduction: let $f_k^y(x) =$ proportion $(Y = y)$ out of k-NN(x)

It estimates $f^y(x) = \mathbb{P}(Y = y|x)$.

... then: $h_k(x) \equiv \mathrm{argmax}_y\{f_k^y(x)\}$, and $h^*(x) = \mathrm{argmax}_y\{f^y(x)\}$

Previous insights extend easily ...

k-NN extends naturally to multiclass



Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y \in \{1, , \ldots, L\}$.

Learn: $h_k(x) =$ majority (Y_i) of k-NN(x).

Reduction: let $f_k^y(x) =$ proportion $(Y = y)$ out of k-NN(x)

It estimates $f^y(x) = \mathbb{P}(Y = y|x)$.

... then: $h_k(x) \equiv \text{argmax}_y\{f_k^y(x)\}$, and $h^*(x) = \text{argmax}_y\{f^y(x)\}$

Previous insights extend easily ...

- **Lipschitzness:** $\|f(x) - f(x')\| \leq \rho(x, x')$

- **Noise margin:** At any x, we want $f^{(1)}(x) \gg f^{(2)}(x)$ ...

$$\text{assume} \quad \mathbb{P}_X\left(f^{(1)}(X) \leq f^{(2)}(X) + \delta\right) \leq \delta^\beta$$

Then: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim (\frac{1}{\log L} \cdot n)^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Same messages as earlier ...

- **Lipschitzness:** $\|f(x) - f(x')\| \leq \rho(x, x')$

- **Noise margin:** At any x, we want $f^{(1)}(x) \gg f^{(2)}(x)$ ...

  assume $\quad \mathbb{P}_X\left(f^{(1)}(X) \leq f^{(2)}(X) + \delta\right) \leq \delta^{\beta}$

Then: $\quad \mathbb{E}\,\mathcal{E}(h_k) \lesssim \left(\frac{1}{\log L} \cdot n\right)^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Same messages as earlier ...

- **Lipschitzness:** $\|f(x) - f(x')\| \leq \rho(x, x')$

- **Noise margin:** At any x, we want $f^{(1)}(x) \gg f^{(2)}(x)$ ...

assume $\quad \mathbb{P}_X\left(f^{(1)}(X) \leq f^{(2)}(X) + \delta\right) \leq \delta^\beta$

Then: $\quad \mathbb{E}\,\mathcal{E}(h_k) \lesssim (\frac{1}{\log L} \cdot n)^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Same messages as earlier ...

- **Lipschitzness:** $\|f(x) - f(x')\| \leq \rho(x, x')$
- **Noise margin:** At any x, we want $f^{(1)}(x) \gg f^{(2)}(x)$ ...

$$\text{assume} \quad \mathbb{P}_X\left(f^{(1)}(X) \leq f^{(2)}(X) + \delta\right) \leq \delta^\beta$$

Then: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim \left(\frac{1}{\log L} \cdot n\right)^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Same messages as earlier ...

- **Lipschitzness:** $\|f(x) - f(x')\| \leq \rho(x, x')$

- **Noise margin:** At any x, we want $f^{(1)}(x) \gg f^{(2)}(x)$ ...

$$\text{assume} \quad \mathbb{P}_X \left(f^{(1)}(X) \leq f^{(2)}(X) + \delta \right) \leq \delta^\beta$$

Then: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim (\frac{1}{\log L} \cdot n)^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Same messages as earlier ...

- **Lipschitzness:** $\|f(x) - f(x')\| \leq \rho(x, x')$

- **Noise margin:** At any x, we want $f^{(1)}(x) \gg f^{(2)}(x)$ ...

$$\text{assume} \quad \mathbb{P}_X \left(f^{(1)}(X) \leq f^{(2)}(X) + \delta \right) \leq \delta^\beta$$

Then: $\mathbb{E}\,\mathcal{E}(h_k) \lesssim (\frac{1}{\log L} \cdot n)^{-(\beta+1)/(2+d)}$, for $k = \Theta(n^{2/(2+d)})$.

Same messages as earlier ...

Mostly open:

Mixed costs regimes (e.g., medicine, finance, ...)

$$y \hookleftarrow \text{Expected cost when } y \text{ is wrong} \neq 1 - \mathbb{P}(Y = y)$$

Natural extensions of previous insights considered in [Reeve, Brown 17]

Practical Question:

assessing mixed costs, and integrating with NN methods ...

Mostly open:

Mixed costs regimes (e.g., medicine, finance, ...)

$y \hookleftarrow$ Expected cost when y is wrong $\neq 1 - \mathbb{P}(Y = y)$

Natural extensions of previous insights considered in [Reeve, Brown 17]

Practical Question:

assessing mixed costs, and integrating with NN methods ...

Mostly open:

Mixed costs regimes (e.g., medicine, finance, ...)

$y \hookleftarrow$ Expected cost when y is wrong $\neq 1 - \mathbb{P}(Y = y)$

Natural extensions of previous insights considered in [Reeve, Brown 17]

Practical Question:

assessing mixed costs, and integrating with NN methods ...

Mostly open:

Mixed costs regimes (e.g., medicine, finance, ...)

$y \hookleftarrow$ Expected cost when y is wrong $\neq 1 - \mathbb{P}(Y = y)$

Natural extensions of previous insights considered in [Reeve, Brown 17]

Practical Question:

assessing mixed costs, and integrating with NN methods ...

Mostly open:

Mixed costs regimes (e.g., medicine, finance, ...)

$$y \hookleftarrow \text{Expected cost when } y \text{ is wrong} \neq 1 - \mathbb{P}(Y = y)$$

Natural extensions of previous insights considered in [Reeve, Brown 17]

Practical Question:

assessing mixed costs, and integrating with NN methods ...

End of Part I