

k -NN Regression adapts to local intrinsic dimension

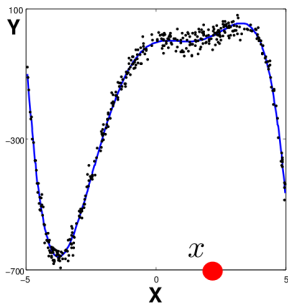
Samory Kpotufe

Max Planck Institute for Intelligent Systems

k-NN Regression

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y = f(X) + \text{noise}$
 $|f(x) - f(x')| \leq \lambda \rho(x, x')$.

Learn: $f_{n,k}(x) = \text{avg}(Y) \text{ of } k\text{-NN}(x)$.

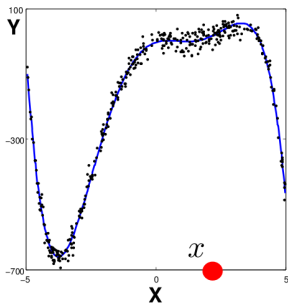


Quite basic! \implies common in practice!

k-NN Regression

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y = f(X) + \text{noise}$
 $|f(x) - f(x')| \leq \lambda \rho(x, x')$.

Learn: $f_{n,k}(x) = \text{avg } (Y) \text{ of } k\text{-NN}(x)$.

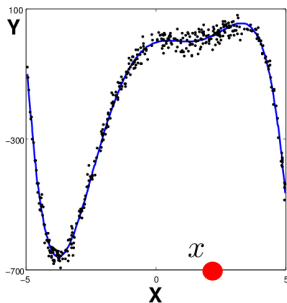


Quite basic! \implies common in practice!

k-NN Regression

Data: $\{(X_i, Y_i)\}_{i=1}^n$, $Y = f(X) + \text{noise}$
 $|f(x) - f(x')| \leq \lambda \rho(x, x')$.

Learn: $f_{n,k}(x) = \text{avg}(Y) \text{ of } k\text{-NN}(x)$.



Quite basic! \implies common in practice!

Curse of dimension: suppose $X \in \mathbb{R}^D$

There exist distributions on (X, Y) such that the *excess risk*

$$|f_{n,k} - f|^2 \doteq \mathbb{E}_x |f_{n,k}(x) - f(x)|^2$$

is of the form $n^{-2/(2+D)}$.

This is true for all nonparametric regressors! 😞 (Stone 82)

Curse of dimension: suppose $X \in \mathbb{R}^D$

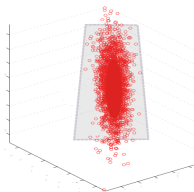
There exist distributions on (X, Y) such that the *excess risk*

$$|f_{n,k} - f|^2 \doteq \mathbb{E}_x |f_{n,k}(x) - f(x)|^2$$

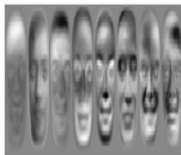
is of the form $n^{-2/(2+D)}$.

This is true for all nonparametric regressors! 😞 (Stone 82)

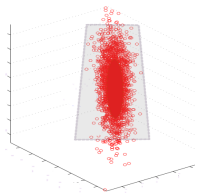
Fortunately, *high* dimensional data often has low intrinsic complexity 😊



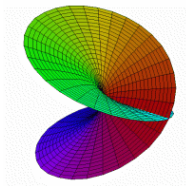
Linear data



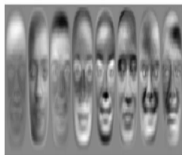
Fortunately, *high* dimensional data often has low intrinsic complexity 😊



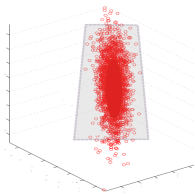
Linear data



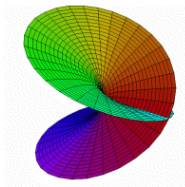
Manifold data



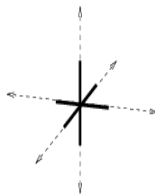
Fortunately, *high* dimensional data often has low intrinsic complexity 😊



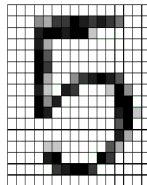
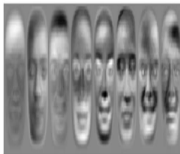
Linear data



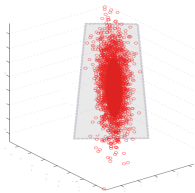
Manifold data



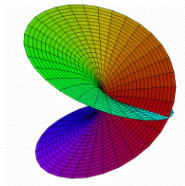
Sparse data



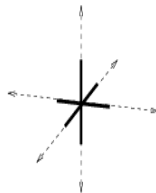
Fortunately, *high* dimensional data often has low intrinsic complexity 😊



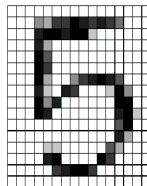
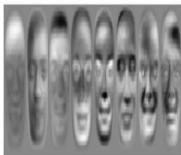
Linear data



Manifold data



Sparse data



Common approach: **Dimension reduction** PCA, Manifold learning (e.g. LLE, Isomap, Laplacian eigenmaps, kernel PCA, ...)

Main result:

k -NN performs well without dimension reduction!

$f_{n,k}$ converges at a rate adaptive to unknown intrinsic dimension.

The result suggests that:

More can be gained tuning k than tuning the parameters of my favorite dimension reduction procedure.

Main result:

k -NN performs well without dimension reduction!

$f_{n,k}$ converges at a rate adaptive to unknown intrinsic dimension.

The result suggests that:

More can be gained tuning k than tuning the parameters of my favorite dimension reduction procedure.

Other work on adaptivity to intrinsic dimension:

- Kernel and local polynomial regression: Bickel and Li 2006, Lafferty and Wasserman 2007.
- Dyadic tree classification: Scott and Nowak 2006.
- 1-NN regression: Kulkarni and Posner 1995.
- RPTree and dyadic tree regression: Kpotufe and Dasgupta 2011.
- Tree-kernel hybrids: Kpotufe 2009.

The above results are under global notions of intrinsic dimension.

Other work on adaptivity to intrinsic dimension:

- Kernel and local polynomial regression: Bickel and Li 2006, Lafferty and Wasserman 2007.
- Dyadic tree classification: Scott and Nowak 2006.
- 1-NN regression: Kulkarni and Posner 1995.
- RPTree and dyadic tree regression: Kpotufe and Dasgupta 2011.
- Tree-kernel hybrids: Kpotufe 2009.

The above results are under global notions of intrinsic dimension.

Outline:

- Intrinsic dimension
- Adaptivity for any $\log n \lesssim k \lesssim n$
- Choosing a good $k = k(x)$

Intrinsic dimension

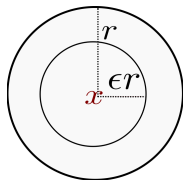


Figure: d -dimensional balls centered at x .

Volume growth: $\text{vol}(B(x, r)) = C \cdot r^d = \epsilon^{-d} \cdot \text{vol}(B(x, \epsilon r))$.

Suppose μ is $\mathcal{U}(B(x, r))$, then $\mu(B(x, r)) \lesssim \epsilon^{-d} \cdot \mu(B(x, \epsilon r))$.

Definition: μ is (C, d) -homogeneous on $B(x, r)$ if $\forall r' \leq r, \epsilon > 0$,
$$\mu(B(x, r')) \leq C \epsilon^{-d} \cdot \mu(B(x, \epsilon r')).$$

Intrinsic dimension

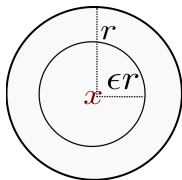


Figure: d -dimensional balls centered at x .

Volume growth: $\text{vol}(B(x, r)) = C \cdot r^d = \epsilon^{-d} \cdot \text{vol}(B(x, \epsilon r))$.

Suppose μ is $\mathcal{U}(B(x, r))$, then $\mu(B(x, r)) \lesssim \epsilon^{-d} \cdot \mu(B(x, \epsilon r))$.

Definition: μ is (C, d) -homogeneous on $B(x, r)$ if $\forall r' \leq r, \epsilon > 0$,
$$\mu(B(x, r')) \leq C \epsilon^{-d} \cdot \mu(B(x, \epsilon r')).$$

Intrinsic dimension

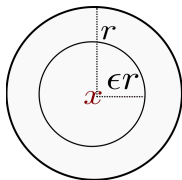


Figure: d -dimensional balls centered at x .

Volume growth: $\text{vol}(B(x, r)) = C \cdot r^d = \epsilon^{-d} \cdot \text{vol}(B(x, \epsilon r))$.

Suppose μ is $\mathcal{U}(B(x, r))$, then $\mu(B(x, r)) \lesssim \epsilon^{-d} \cdot \mu(B(x, \epsilon r))$.

Definition: μ is (C, d) -homogeneous on $B(x, r)$ if $\forall r' \leq r, \epsilon > 0$,
$$\mu(B(x, r')) \leq C \epsilon^{-d} \cdot \mu(B(x, \epsilon r')).$$

Intrinsic dimension

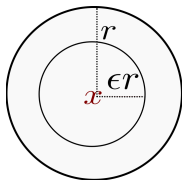


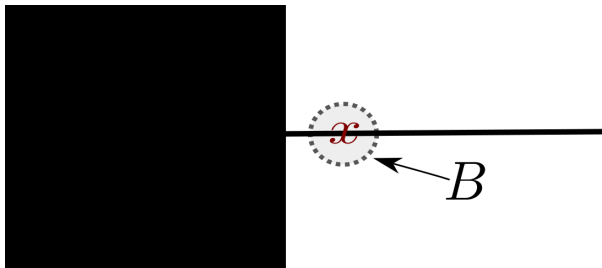
Figure: d -dimensional balls centered at x .

Volume growth: $\text{vol}(B(x, r)) = C \cdot r^d = \epsilon^{-d} \cdot \text{vol}(B(x, \epsilon r))$.

Suppose μ is $\mathcal{U}(B(x, r))$, then $\mu(B(x, r)) \lesssim \epsilon^{-d} \cdot \mu(B(x, \epsilon r))$.

Definition: μ is (C, d) -homogeneous on $B(x, r)$ if $\forall r' \leq r, \epsilon > 0$,
$$\mu(B(x, r')) \leq C \epsilon^{-d} \cdot \mu(B(x, \epsilon r')).$$

Given a query x , the behavior of μ in a neighborhood B of x can capture the intrinsic dimension in B .

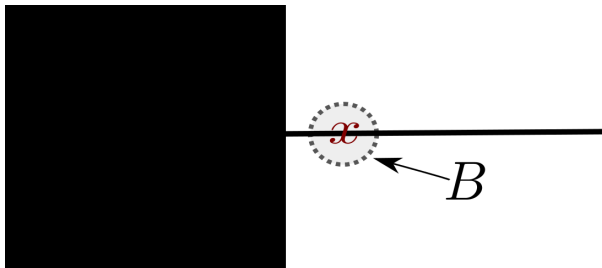


Location of query x matters!

Size of neighborhood B matters!

For k -NN, size of relevant neighborhood B will depend on k and n .

Given a query x , the behavior of μ in a neighborhood B of x can capture the intrinsic dimension in B .

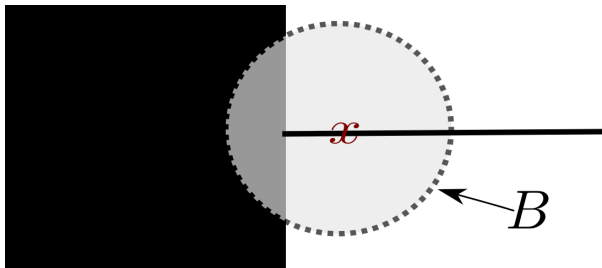


Location of query x matters!

Size of neighborhood B matters!

For k -NN, size of relevant neighborhood B will depend on k and n .

Given a query x , the behavior of μ in a neighborhood B of x can capture the intrinsic dimension in B .

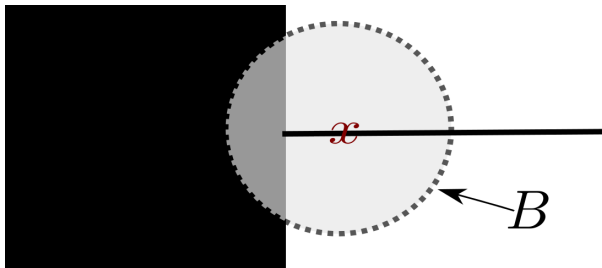


Location of query x matters!

Size of neighborhood B matters!

For k -NN, size of relevant neighborhood B will depend on k and n .

Given a query x , the behavior of μ in a neighborhood B of x can capture the intrinsic dimension in B .



Location of query x matters!

Size of neighborhood B matters!

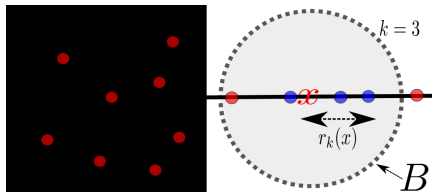
For k -NN, size of relevant neighborhood B will depend on k and n .

Outline:

- Intrinsic dimension
- Adaptivity for any $\log n \lesssim k \lesssim n$
- Choosing a good $k = k(x)$

Adaptivity for k - General intuition:

Fix, $n \gtrsim k \gtrsim \log n$, and let $x \in \text{region } B$ of dimension d .



Rate of convergence of $f_{n,k}(x)$ depends on:

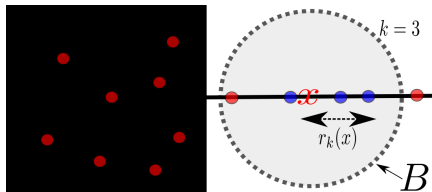
- (**Variance** of $f_{n,k}(x)$) $\approx 1/k$.
- (**Bias** of $f_{n,k}(x)$) $\approx r_k(x)$.

It turns out: $r_k(x) \approx (k/n)^{1/d}$

Also: $r_k(x)$ depends on $\mu(B)$ (smaller in sparse regions)

Adaptivity for k - General intuition:

Fix, $n \gtrsim k \gtrsim \log n$, and let $x \in \text{region } B$ of dimension d .



Rate of convergence of $f_{n,k}(x)$ depends on:

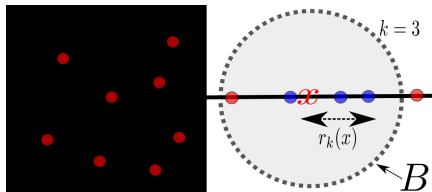
- (**Variance** of $f_{n,k}(x)$) $\approx 1/k$.
- (**Bias** of $f_{n,k}(x)$) $\approx r_k(x)$.

It turns out: $r_k(x) \approx (k/n)^{1/d}$.

Also: $r_k(x)$ depends on $\mu(B)$ (smaller in dense regions B).

Adaptivity for k - General intuition:

Fix, $n \gtrsim k \gtrsim \log n$, and let $x \in \text{region } B$ of dimension d .



Rate of convergence of $f_{n,k}(x)$ depends on:

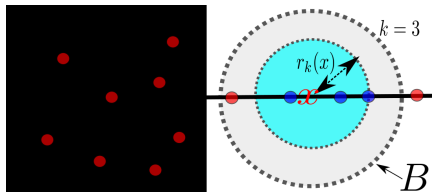
- (**Variance** of $f_{n,k}(x)$) $\approx 1/k$.
- (**Bias** of $f_{n,k}(x)$) $\approx r_k(x)$.

It turns out: $r_k(x) \approx (k/n)^{1/d}$.

Also: $r_k(x)$ depends on $\mu(B)$ (smaller in dense regions B).

Adaptivity for k - General intuition:

Fix, $n \gtrsim k \gtrsim \log n$, and let $x \in \text{region } B$ of dimension d .



Rate of convergence of $f_{n,k}(x)$ depends on:

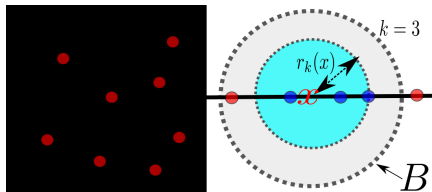
- (**Variance** of $f_{n,k}(x)$) $\approx 1/k$.
- (**Bias** of $f_{n,k}(x)$) $\approx r_k(x)$.

It turns out: $r_k(x) \approx (k/n)^{1/d}$.

Also: $r_k(x)$ depends on $\mu(B)$ (smaller in dense regions B).

Adaptivity for k - General intuition:

Fix, $n \gtrsim k \gtrsim \log n$, and let $x \in \text{region } B$ of dimension d .



Rate of convergence of $f_{n,k}(x)$ depends on:

- (**Variance** of $f_{n,k}(x)$) $\approx 1/k$.
- (**Bias** of $f_{n,k}(x)$) $\approx r_k(x)$.

It turns out: $r_k(x) \approx (k/n)^{1/d}$.

Also: $r_k(x)$ depends on $\mu(B)$ (smaller in dense regions B).

Adaptivity for k - Result:

Theorem: The following holds w.h.p. simultaneously for all $x \in \mathcal{X}$ and $\log n \lesssim k \lesssim n$.

Consider any B centered at x , s.t. $\mu(B) \gtrsim k/n$. Suppose μ is (C, d) -homogeneous on B . We have

$$|f_{n,k}(x) - f(x)|^2 \lesssim \frac{1}{k} + \lambda^2 \left(\frac{Ck}{n\mu(B)} \right)^{1/d}.$$

Rate is best if x is in a dense region B with low dimension d .

Adaptivity for k - Result:

Theorem: The following holds w.h.p. simultaneously for all $x \in \mathcal{X}$ and $\log n \lesssim k \lesssim n$.

Consider any B centered at x , s.t. $\mu(B) \gtrsim k/n$. Suppose μ is (C, d) -homogeneous on B . We have

$$|f_{n,k}(x) - f(x)|^2 \lesssim \frac{1}{k} + \lambda^2 \left(\frac{Ck}{n\mu(B)} \right)^{1/d}.$$

Rate is best if x is in a dense region B with low dimension d .

Outline:

- Intrinsic dimension
- Adaptivity for any $\log n \lesssim k \lesssim n$
- Choosing a good $k = k(x)$

Choosing $k(x)$ - Best possible rate in terms of d

Theorem: Consider a metric measure space (\mathcal{X}, ρ, μ) , such that for all $x \in \mathcal{X}, r > 0, \epsilon > 0$, we have $\mu(B(x, r)) \approx \epsilon^{-d} \mu(B(x, \epsilon r))$. Then, for any regressor f_n , there exists $\mathcal{D}_{X,Y}$ with marginal μ and where $f(x) = \mathbb{E} Y|x$ is λ -Lipschitz, such that

$$\mathbb{E}_{S \sim \mathcal{D}_{X,Y}^n, X \sim \mu} |f_n(X) - f(X)|^2 \gtrsim \lambda^{2d/(2+d)} \cdot n^{-2/(2+d)}.$$

Choosing $k(x)$ - Best possible rate in terms of d

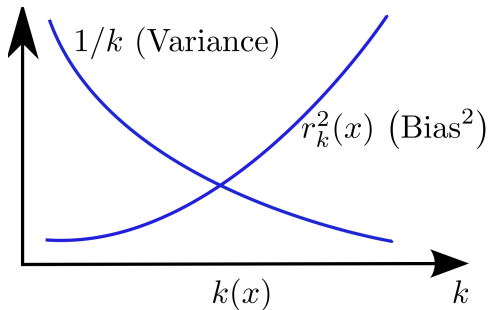
Theorem: Consider a metric measure space (\mathcal{X}, ρ, μ) , such that for all $x \in \mathcal{X}, r > 0, \epsilon > 0$, we have $\mu(B(x, r)) \approx \epsilon^{-d} \mu(B(x, \epsilon r))$. Then, for any regressor f_n , there exists $\mathcal{D}_{X,Y}$ with marginal μ and where $f(x) = \mathbb{E} Y|x$ is λ -Lipschitz, such that

$$\mathbb{E}_{S \sim \mathcal{D}_{X,Y}^n, X \sim \mu} |f_n(X) - f(X)|^2 \gtrsim \lambda^{2d/(2+d)} \cdot n^{-2/(2+d)}.$$

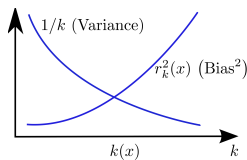
Choosing k locally at x - Intuition

Note: Cross-validation and dimension estimation require large samples sizes, which is unlikely in small neighborhoods of x .

Instead:



Choosing $k(x)$ - Result



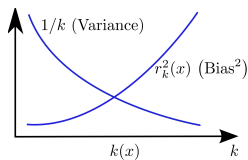
Theorem: Suppose $k(x)$ is chosen as above. The following holds w.h.p. simultaneously for all x .

Consider any B centered at x , s.t. $\mu(B) \gtrsim n^{-1/3}$. Suppose μ is (C, d) -homogeneous on B . We have

$$|f_{n,k}(x) - f(x)|^2 \lesssim \lambda^2 \left(\frac{C}{n\mu(B)} \right)^{2/(2+d)}.$$

As $n \rightarrow \infty$ the claim applies to any B centered at x , $\mu(B) \neq 0$.

Choosing $k(x)$ - Result



Theorem: Suppose $k(x)$ is chosen as above. The following holds w.h.p. simultaneously for all x .

Consider any B centered at x , s.t. $\mu(B) \gtrsim n^{-1/3}$. Suppose μ is (C, d) -homogeneous on B . We have

$$|f_{n,k}(x) - f(x)|^2 \lesssim \lambda^2 \left(\frac{C}{n\mu(B)} \right)^{2/(2+d)}.$$

As $n \rightarrow \infty$ the claim applies to any B centered at x , $\mu(B) \neq 0$.

Results likely extend to:

- Higher order polynomial regression/classification using k -NN.
- Local choice of bandwidth in kernel regression.

Take home message

k -NN regression performs well without dimension reduction!

Question:

Is there a general principle for designing adaptive learners?

Thank you for listening.

Take home message

k -NN regression performs well without dimension reduction!

Question:

Is there a general principle for designing adaptive learners?

Thank you for listening.

Take home message

k -NN regression performs well without dimension reduction!

Question:

Is there a general principle for designing adaptive learners?

Thank you for listening.

Take home message

k -NN regression performs well without dimension reduction!

Question:

Is there a general principle for designing adaptive learners?

Thank you for listening.

PS: looking for a job 😊