Nearest Neighbors I: Regression and Classification

Kamalika Chaudhuri

University of California, San Diego

Talk Outline

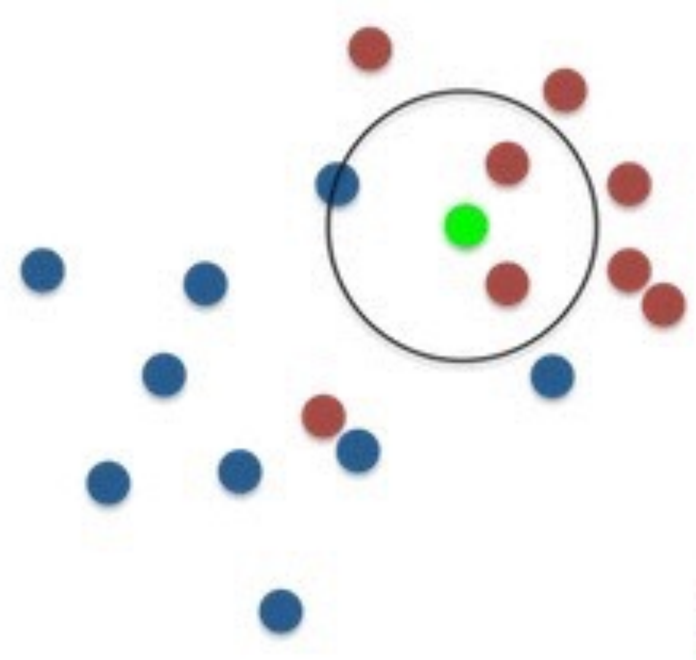
- Part I: k-Nearest neighbors: Regression and Classification
- Part II: k-Nearest neighbors (and other non-parametrics): Adversarial examples

k Nearest Neighbors

Given: training data $(x_1, y_1), \ldots, (x_n, y_n)$ in X x $\{0, 1\}$

query point x

Predict y for x from the k closest neighbors of x among x_i



Example:

k-NN classification: predict majority label of k closest neighbors

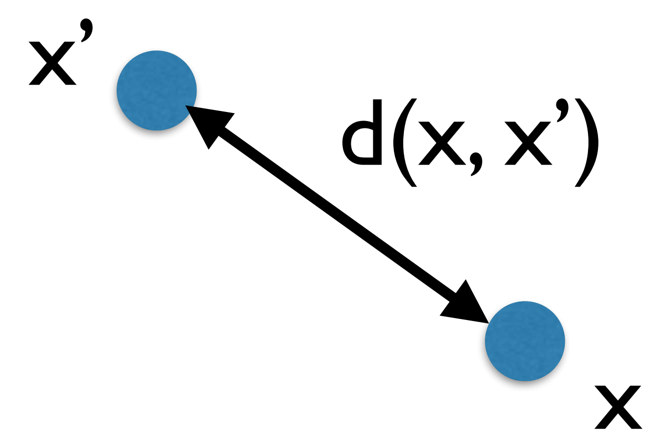
k-NN regression: predict average label of k closest neighbors

The Metric Space

Data points lie in metric space with distance function d

Examples:

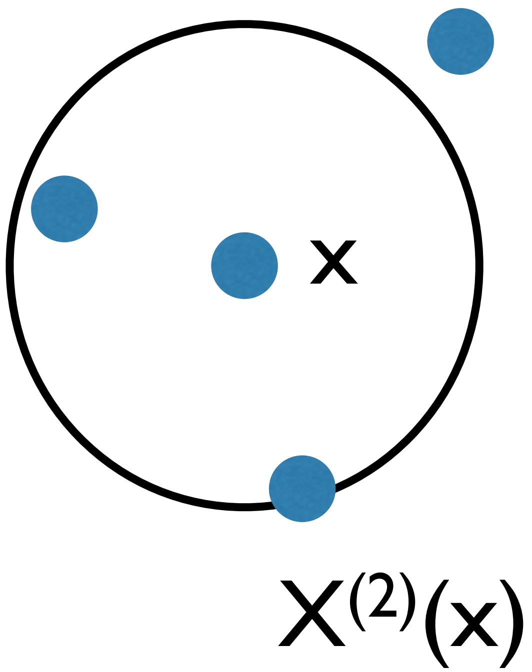
- $X = R^D$, d = Euclidean distance
- $X = R^D$, d = l_p distance
- Metric based on user preferences



Notation

$X^{(i)}(x)$ = i-th nearest neighbor of x

$Y^{(i)}(x)$ = label of $X^{(i)}(x)$



Tutorial Outline

- Nearest Neighbor Regression
  - The Setting
  - Universal Consistency
  - Rates of Convergence
- Nearest Neighbor Classification
  - The Statistical Learning Framework
  - Consistency
  - Rates of Convergence

NN Regression Setting

Compact metric space (X, d)

Uniform measure μ on X (for now)

NN Regression Setting

Compact metric space (X, d)

Uniform measure μ on X (for now)

Input: Training data $(x_1, y_1), \ldots, (x_n, y_n)$

where: $x_i \sim \mu$

$$y_i = f(x_i) + noise$$

unknown f

NN Regression Setting

Compact metric space (X, d)

Uniform measure μ on X (for now)

Input: Training data $(x_1, y_1), \ldots, (x_n, y_n)$

where: $x_i \sim \mu$

$$y_i = f(x_i) + noise$$

unknown f

k-NN Regressor: $$\hat{f}_k(x) = \frac{1}{k} \sum_{i=1}^{k} Y^{(i)}(x)$$

Universality

Input: Training data $(x_1, y_1), \ldots, (x_n, y_n)$

where $y_i = f(x_i) + noise$

k-NN Regressor: $\hat{f}_k(x) = \frac{1}{k} \sum_{i=1}^{k} Y^{(i)}(x)$

What f can k-NN regression represent?

Universality

Input: Training data $(x_1, y_1), \ldots, (x_n, y_n)$

where $y_i = f(x_i) + noise$

k-NN Regressor: $\hat{f}_k(x) = \frac{1}{k}\sum_{i=1}^{k} Y^{(i)}(x)$

What f can k-NN regression represent?

Answer: Any f, provided k grows suitably with n [Devroye, Gyorfi, Kryzak, Lugosi, 94]

More Formally…

k_n NN Regression: when k grows with n

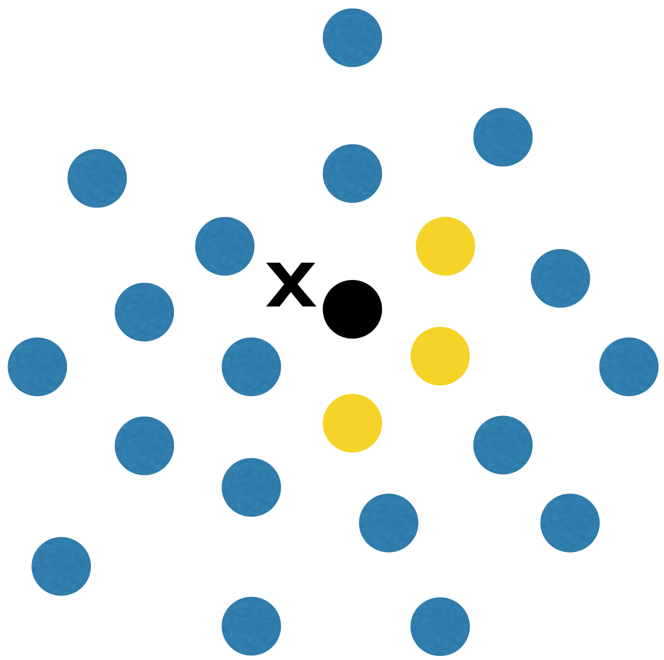
Theorem: If $k_n \to \infty$ and if $k_n/n \to 0$, then for any f,

$$\mathbb{E}_{X\sim\mu}[|f(X) - \hat{f}_{k_n}(X)|] \to 0$$

as $n \to \infty$

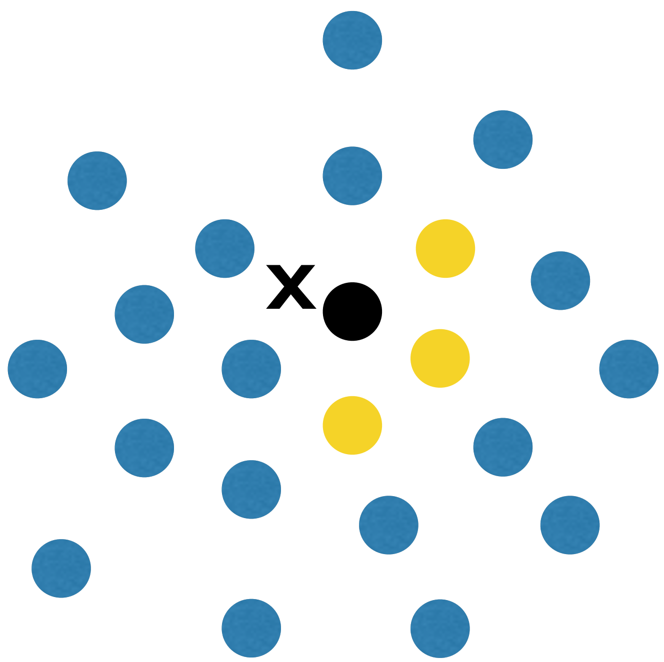
k_n NN Regression is **universally consistent**

Intuition: Universal Consistency



As n grows, $X^{(i)}(x)$ move closer to x (continuous μ)

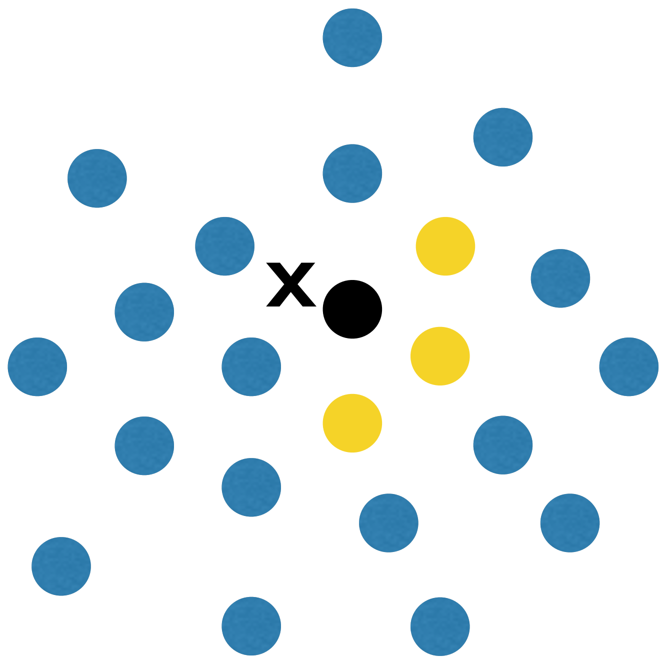
Intuition: Universal Consistency



As n grows, $X^{(i)}(x)$ move closer to x (continuous μ)

If k_n is constant or grows slowly $(k_n/n \to 0)$ then $X^{(i)}(x) \to x, i \leq k_n$

Intuition: Universal Consistency

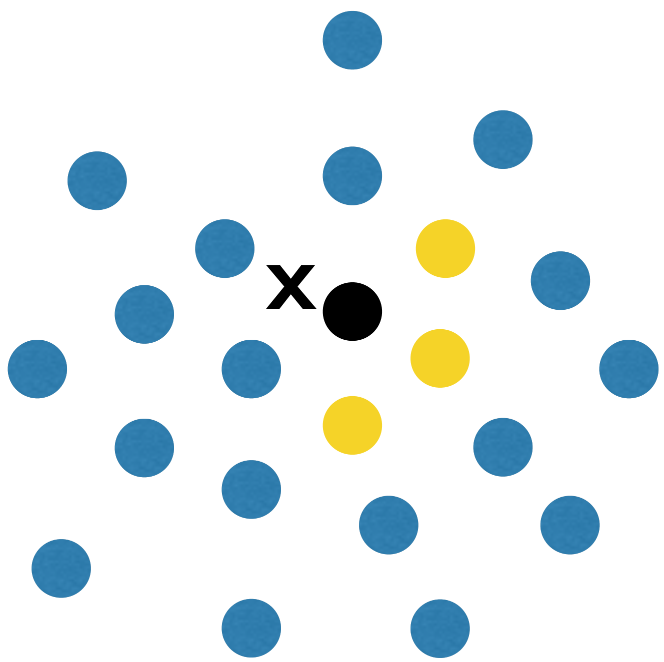


As n grows, $X^{(i)}(x)$ move closer to x (continuous μ)

If k_n is constant or grows slowly $(k_n/n \to 0)$ then $X^{(i)}(x) \to x, i \leq k_n$

If f is continuous, then $f(X^{(i)}(x)) \to f(x), 1 \leq i \leq k_n$

Intuition: Universal Consistency



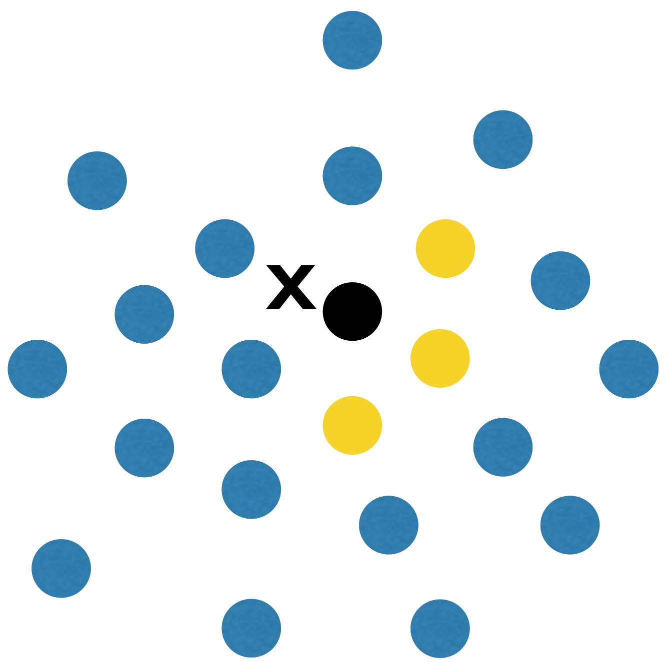
As n grows, $X^{(i)}(x)$ move closer to x (continuous μ)

If k_n is constant or grows slowly $(k_n/n \to 0)$ then $X^{(i)}(x) \to x, i \leq k_n$

If f is continuous, then $f(X^{(i)}(x)) \to f(x), 1 \leq i \leq k_n$

If $k_n \to \infty$, then $$\frac{1}{k_n}\sum_{i=1}^{k_n}(f(X^{(i)}(x)) + \text{noise}) \to f(x)$$

Intuition: Universal Consistency



As n grows, $X^{(i)}(x)$ move closer to x (continuous μ)

If k_n is constant or grows slowly $(k_n/n \to 0)$ then $X^{(i)}(x) \to x, i \leq k_n$

If f is continuous, then $f(X^{(i)}(x)) \to f(x), 1 \leq i \leq k_n$

If $k_n \to \infty$, then $\frac{1}{k_n}\sum_{i=1}^{k_n}(f(X^{(i)}(x)) + \text{noise}) \to f(x)$

Any f can be approximated arbitrarily well by continuous f

$\blacksquare$

Tutorial Outline

- Nearest Neighbor Regression
  - The Setting
  - Universality
  - Rates of Convergence
- Nearest Neighbor Classification
  - The Statistical Learning Framework
  - Consistency
  - Rates of Convergence

Convergence Rates

Definition: f is L-Lipschitz if for all x and x',

$$|f(x) - f(x')| \leq L \cdot d(x, x')$$

Convergence Rates

Definition: f is L-Lipschitz if for all x and x',

$$|f(x) - f(x')| \leq L \cdot d(x, x')$$

Theorem: If f is L-Lipschitz then for $k_n = \Theta(n^{2/(2+D)})$, there exists a constant C such that

$$\mathbb{E}_{x \sim \mu}[\|\hat{f}_k(x) - f(x)\|^2] \leq Cn^{-2/(2+D)}$$ (D = data dim)

Convergence Rates

Definition: f is L-Lipschitz if for all x and x',

$$|f(x) - f(x')| \leq L \cdot d(x, x')$$

Theorem: If f is L-Lipschitz then for $k_n = \Theta(n^{2/(2+D)})$, there exists a constant C such that

$$\mathbb{E}_{x \sim \mu}[\|\hat{f}_k(x) - f(x)\|^2] \leq Cn^{-2/(2+D)}$$ (D = data dim)

Better bounds for low intrinsic dimension [Kpotufe11]

$k_n = \Theta(n^{2/(2+D)})$ is the optimal value of k_n

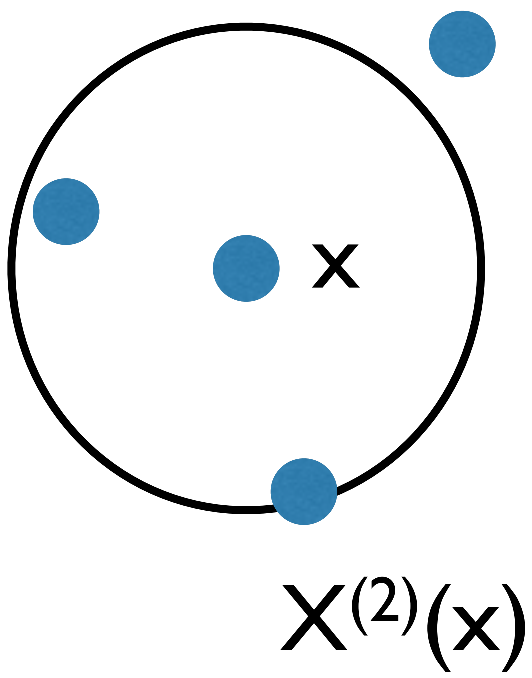
How fast is convergence?

- How small are k-NN distances?
- From distances to convergence rates

k-NN Distances

Given i.i.d. $x_1, \ldots, x_n \sim \mu$

Define: $r_k(x) = d(x, X^{(k)}(x))$

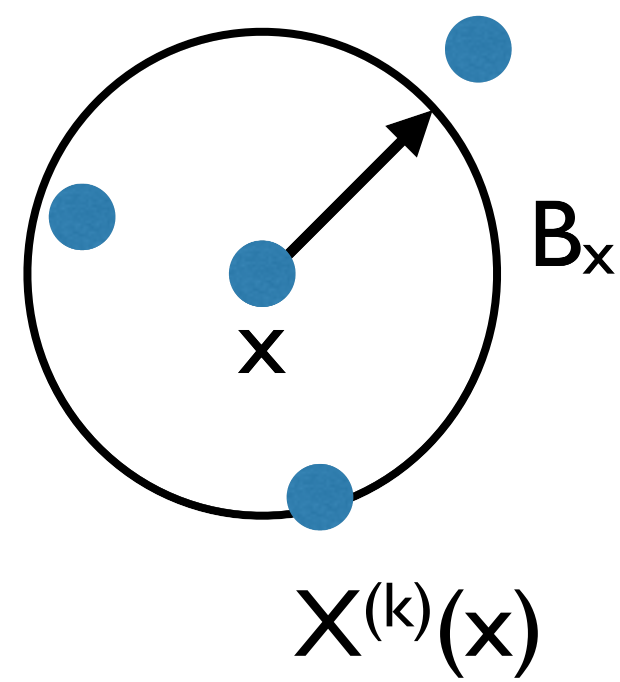


How small is $r_k(x)$?

k-NN Distances

Given i.i.d. $x_1, \ldots, x_n \sim \mu$

Define: $r_k(x) = d(x, X^{(k)}(x))$

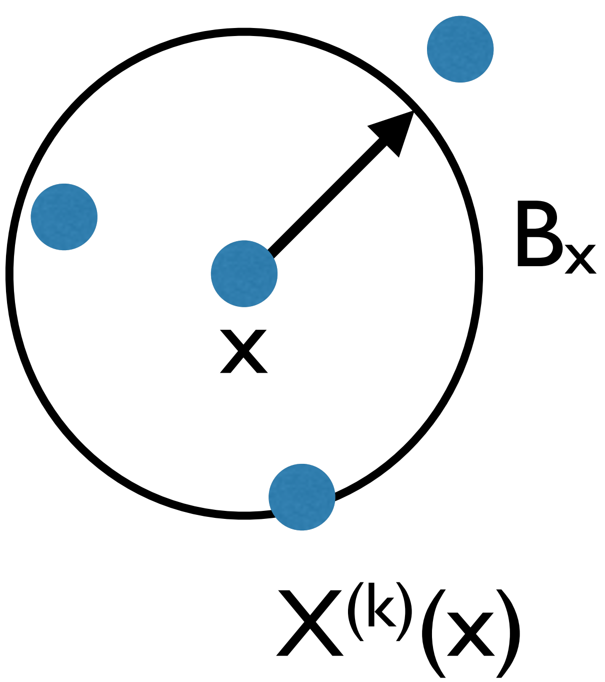


k-NN Distances

Given i.i.d. $x_1, \ldots, x_n \sim \mu$

Define: $r_k(x) = d(x, X^{(k)}(x))$

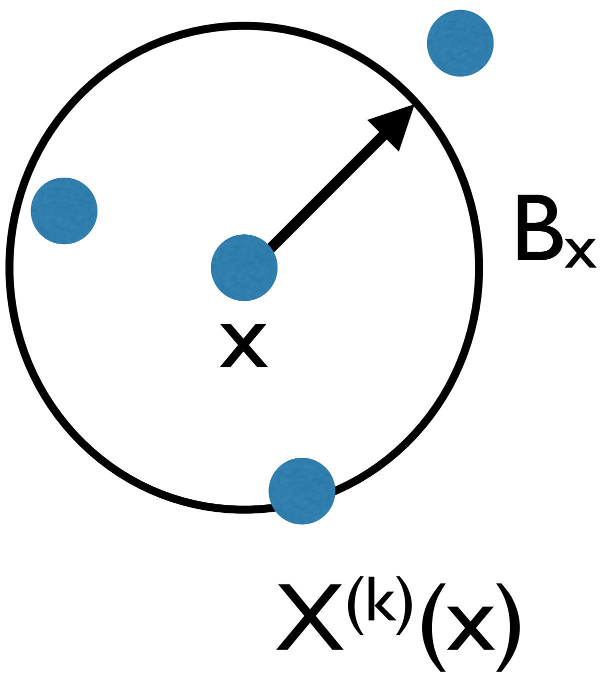
Let B_x = Ball$(x, r_k(x))$



k-NN Distances

Given i.i.d. $x_1, \ldots, x_n \sim \mu$

Define: $r_k(x) = d(x, X^{(k)}(x))$



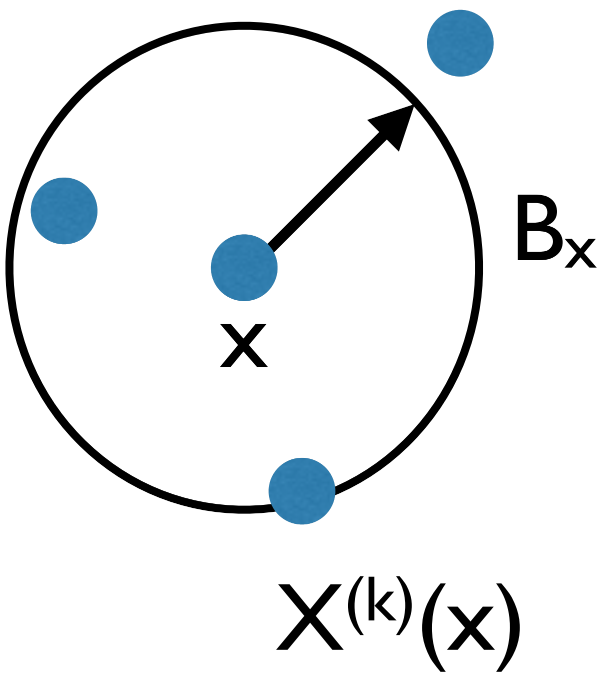
Let B_x = Ball(x, r_k(x))

$\hat{\mu}(B_x) = k/n \quad \approx \mu(B_x)$ (whp for large k, n)

k-NN Distances

Given i.i.d. $x_1, \ldots, x_n \sim \mu$

Define: $r_k(x) = d(x, X^{(k)}(x))$



Let B_x = Ball(x, r_k(x))

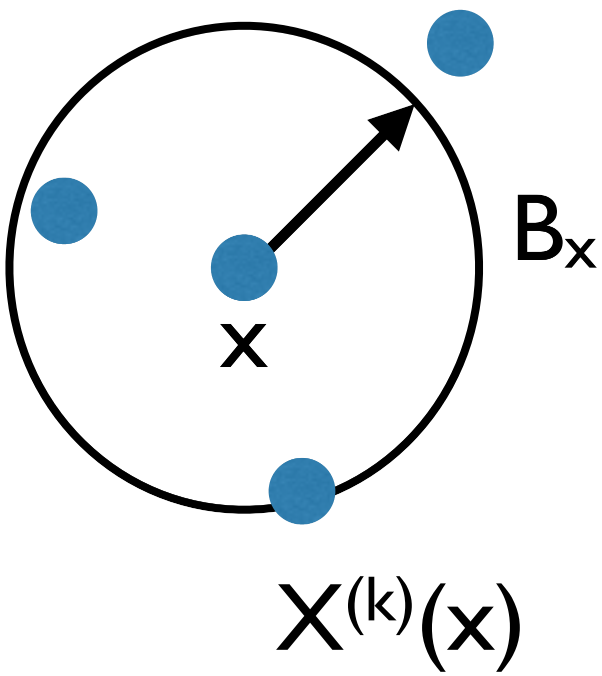
$\hat{\mu}(B_x) = k/n \quad \approx \mu(B_x)$ (whp for large k, n)

$$\mu(B_x) = \int_{B_x} \mu(x')dx' \approx \mu(x) \int_{B_x} dx' \approx \mu(x) r_k(x)^D$$

k-NN Distances

Given i.i.d. $x_1, \ldots, x_n \sim \mu$

Define: $r_k(x) = d(x, X^{(k)}(x))$



Let B_x = Ball(x, r_k(x))

$\hat{\mu}(B_x) = k/n \quad \approx \mu(B_x)$ (whp for large k, n)

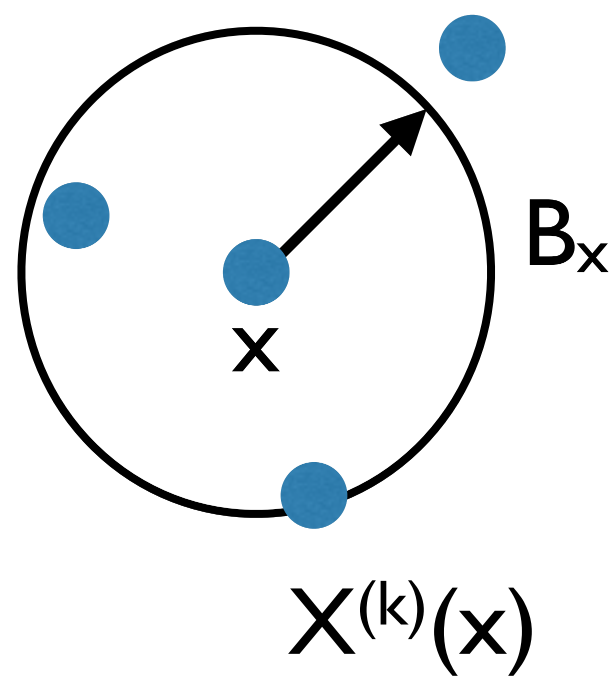
$$\mu(B_x) = \int_{B_x} \mu(x')dx' \approx \mu(x) \int_{B_x} dx' \approx \mu(x) r_k(x)^D$$

$$r_k(x) \approx \left(\frac{1}{\mu(x)} \cdot \frac{k}{n} \right)^{1/D}$$ (D = data dimension)

k-NN Distances

Given i.i.d. $x_1, \ldots, x_n \sim \mu$

Define: $r_k(x) = d(x, X^{(k)}(x))$



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

(Curse of dimensionality)

Better for data with low intrinsic dimension

[Kpotufe, 2011], [Samworth 12], [Costa and Hero 04]

From Distances to Rates

1. Bias-Variance Decomposition

2. Bound Bias and Variance in terms of distances

3. Integrate over the space

Bias-Variance Decomposition

For a fixed x, and {x_i}, define:

$$\tilde{f}_k(x) = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[Y^{(i)}(x)|\{x_i\}]$$

Then:

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

Bounding Bias and Variance

Bounding bias: For any x,

$$\|\tilde{f}_k(x) - f(x)\|^2 \leq \left(\frac{1}{k} \sum_{i=1}^{k} |f(x) - f(X^{(i)}(x)| \right)^2$$

Bounding Bias and Variance

Bounding bias: For any x,

$$\|\tilde{f}_k(x) - f(x)\|^2 \leq \left(\frac{1}{k} \sum_{i=1}^{k} |f(x) - f(X^{(i)}(x)| \right)^2$$

$$\leq (L \cdot d(x, X^{(k)}(x)))^2 \qquad \text{(by Lipschitzness)}$$

Bounding Bias and Variance

Bounding bias: For any x,

$$\|\tilde{f}_k(x) - f(x)\|^2 \leq \left(\frac{1}{k} \sum_{i=1}^{k} |f(x) - f(X^{(i)}(x)| \right)^2$$

$$\leq (L \cdot d(x, X^{(k)}(x)))^2 \qquad \text{(by Lipschitzness)}$$

$$\leq \Theta \left(\frac{k}{n} \right)^{2/D} \qquad \text{(from distances)}$$

Bounding Bias and Variance

Bounding bias: For any x,

$$\|\tilde{f}_k(x) - f(x)\|^2 \leq \left(\frac{1}{k} \sum_{i=1}^{k} |f(x) - f(X^{(i)}(x)| \right)^2$$

$$\leq (L \cdot d(x, X^{(k)}(x)))^2 \qquad \text{(by Lipschitzness)}$$

$$\leq \Theta \left(\frac{k}{n} \right)^{2/D} \qquad \text{(from distances)}$$

Bounding variance:

$$\mathbb{E}[\|f_k(x) - \tilde{f}_k(x)\|^2] = \mathbb{E}\left(\frac{1}{k}(Y^{(i)}(x) - \mathbb{E}[Y^{(i)}(x)])^2 \right) = \frac{\sigma_Y^2}{k}$$

Integrating across the space

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

$$\lessapprox \frac{1}{k} + \left(\frac{k}{n}\right)^{2/D}$$

Integrating across the space

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

$$\lessapprox \frac{1}{k} + \left(\frac{k}{n}\right)^{2/D}$$

Optimizing for k: $k_n = \Theta(n^{2/(2+D)})$

$\blacksquare$

Integrating across the space

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

$$\lessapprox \frac{1}{k} + \left(\frac{k}{n}\right)^{2/D}$$

Optimizing for k: $k_n = \Theta(n^{2/(2+D)})$

Which leads to: $\mathbb{E}[\|f_k(x) - f(x)\|^2] \leq n^{-2/(2+D)}$

Bound is optimal, better for low intrinsic dimension ■

Tutorial Outline

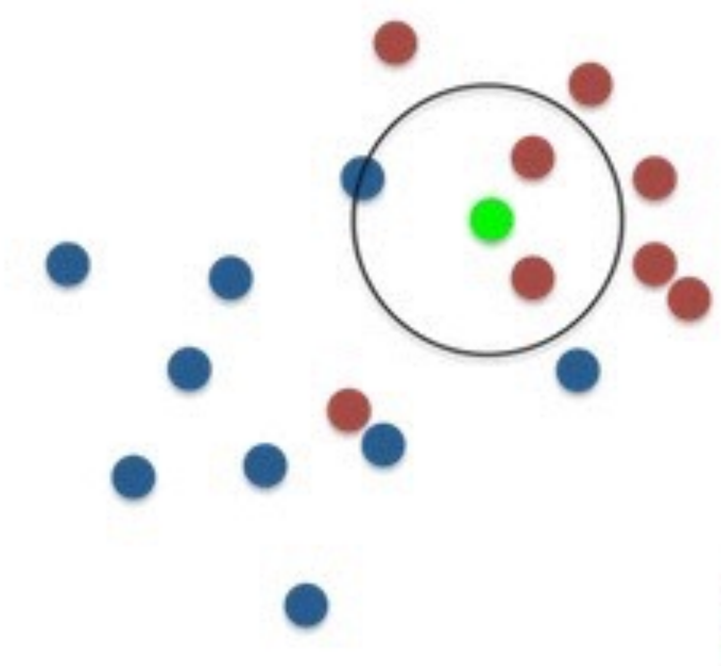
- Nearest Neighbor Regression
  - The Setting
  - Universality
  - Rates of Convergence
- Nearest Neighbor Classification
  - The Statistical Learning Framework
  - Consistency
  - Rates of Convergence

Nearest Neighbor Classification

Given: training data $(x_1, y_1), \ldots, (x_n, y_n)$ in X x {0, 1}

query point x

Predict majority label of the k closest points closest to x



$h_{n,k}$ = k-NN classifier on n points

$h_{n,k}(x) = 0$, if $\frac{1}{k}\sum_{i=1}^{k} Y^{(i)}(x) \leq \frac{1}{2}$

= 1, otherwise

The Statistical Learning Framework

Metric space (X, d)

Underlying measure μ on X from which points are drawn

Label of x is a coin flip with bias $\eta(x) = \Pr(y = 1|x)$

Risk or error of a classifier h: $R(h) = \Pr(h(X) \neq Y)$

Accuracy(h) = 1 - R(h)

Goal: Find h that minimizes risk or maximizes accuracy

The Bayes Optimal Classifier

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

Risk(h) = $\mathbb{E}_X[\min(\eta(X), 1 - \eta(X))]$ = R*

The Bayes Optimal Classifier minimizes risk

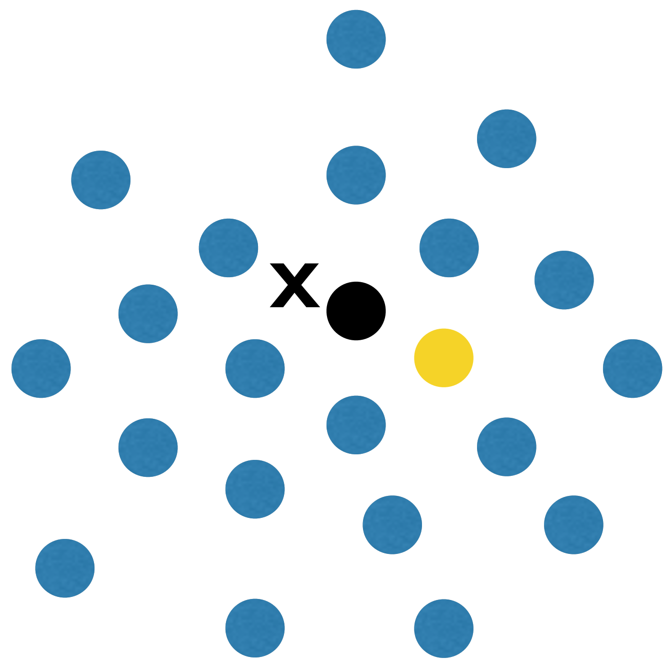
Tutorial Outline

- Nearest Neighbor Regression
  - The Setting
  - Universality
  - Rates of Convergence
- Nearest Neighbor Classification
  - The Statistical Learning Framework
  - Consistency
  - Rates of Convergence

Consistency

Does $R(h_{n,k})$ converge to R^* as n goes to infinity?

Consistency of 1-NN

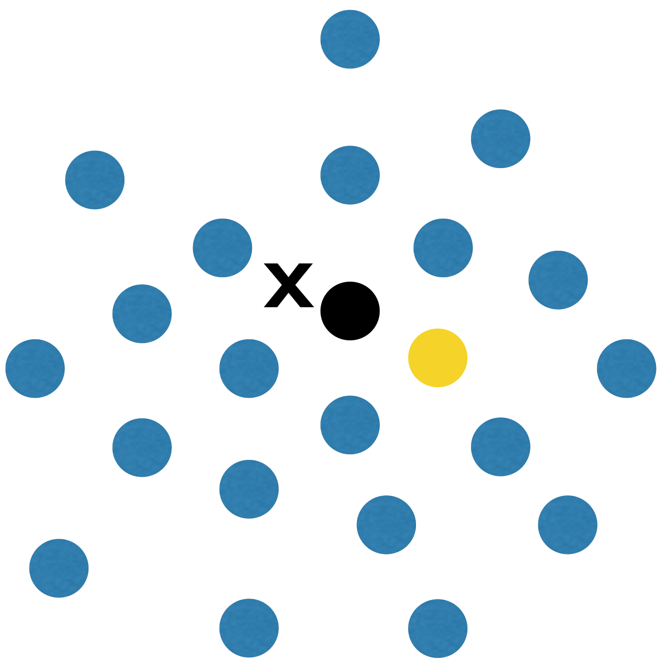


Assume:

Continous η

Absolutely continuous μ

Consistency of 1-NN



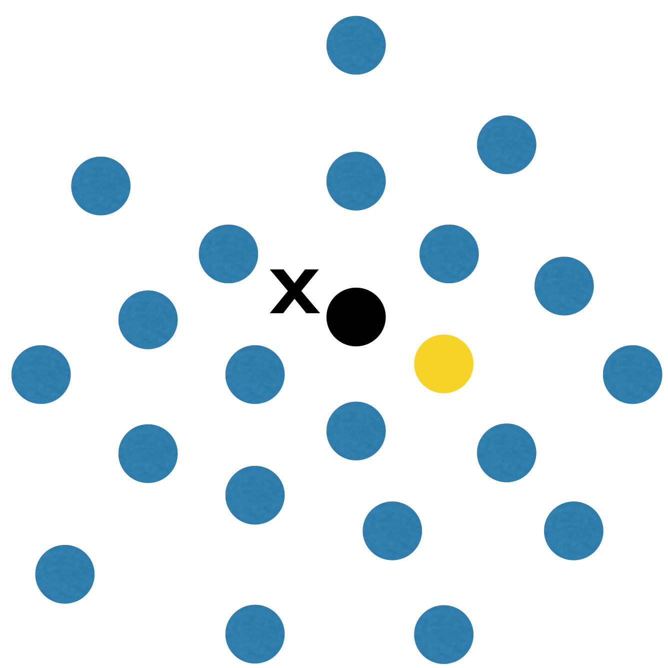
Assume:

Continous η

Absolutely continuous μ

$$R(h_{n,1}) \to \mathbb{E}_X[2\eta(X)(1-\eta(X))] \neq R^*$$

Consistency of 1-NN



Assume:

Continous η

Absolutely continuous μ

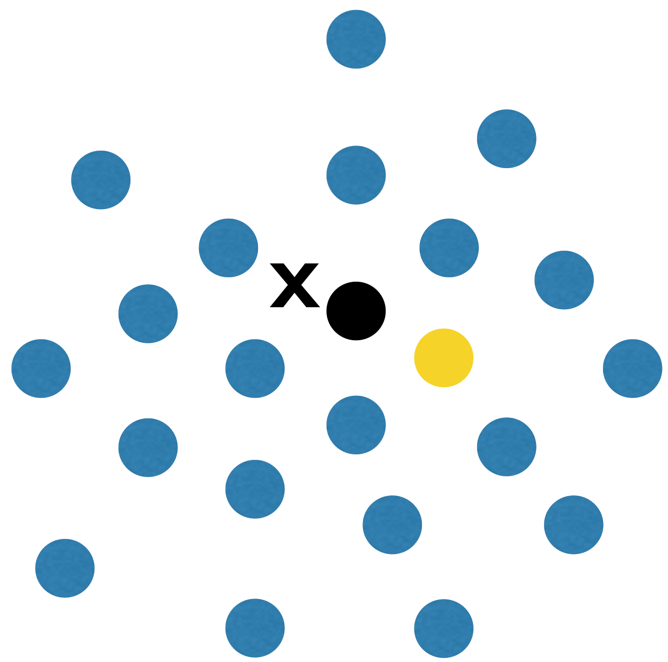
$$R(h_{n,1}) \to \mathbb{E}_X[2\eta(X)(1-\eta(X))] \neq R^*$$

1-NN is inconsistent

k-NN for constant k is also inconsistent

[Cover and Hart, 67]

Proof Intuition



Assume:

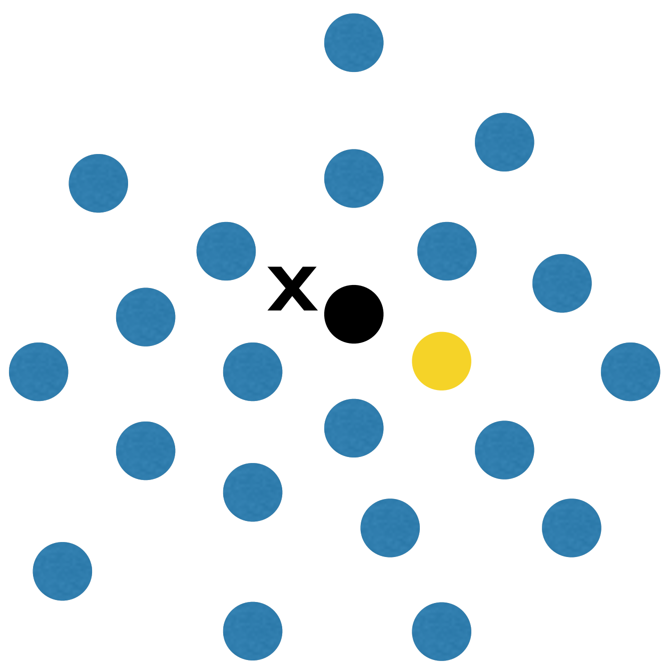
Continous η

Absolutely continuous μ

For any x, $X^{(1)}(x)$ converges to x

$\blacksquare$

Proof Intuition



Assume:

Continous η

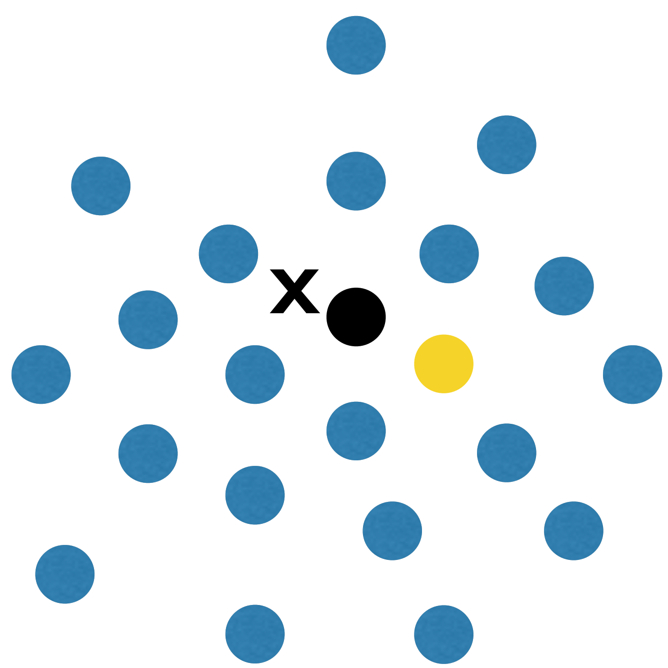
Absolutely continuous μ

For any x, $X^{(1)}(x)$ converges to x

By continuity, $\eta(X^{(1)}(x)) \to \eta(x)$

$\blacksquare$

Proof Intuition



Assume:

Continous η

Absolutely continuous μ

For any x, $X^{(1)}(x)$ converges to x

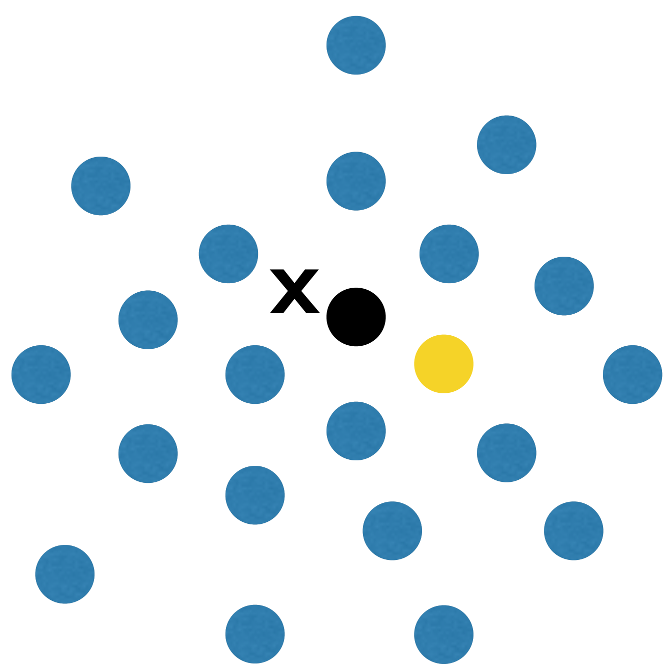
By continuity, $\eta(X^{(1)}(x)) \to \eta(x)$

$$\Pr(Y^{(1)}(x) \neq y) = \eta(x)(1 - \eta(X^{(1)}(x)) + \eta(X^{(1)}(x))(1 - \eta(x))$$

$$\to 2\eta(x)(1 - \eta(x))$$

$\blacksquare$

Proof Intuition



Assume:

Continous η

Absolutely continuous μ

For any x, $X^{(1)}(x)$ converges to x

By continuity, $\eta(X^{(1)}(x)) \to \eta(x)$

$$\Pr(Y^{(1)}(x) \neq y) = \eta(x)(1 - \eta(X^{(1)}(x)) + \eta(X^{(1)}(x))(1 - \eta(x))$$
$$\to 2\eta(x)(1 - \eta(x))$$

Thus: $R(h_{n,1}) \to \mathbb{E}_X[2\eta(X)(1 - \eta(X))] \neq R^*$ ∎

Consistency under Continuity

Assume η is continuous

Theorem: If $k_n \to \infty$ and if $k_n/n \to 0$, then

$$R(h_{n,k_n}) \to R^* \quad \text{as} \quad n \to \infty$$

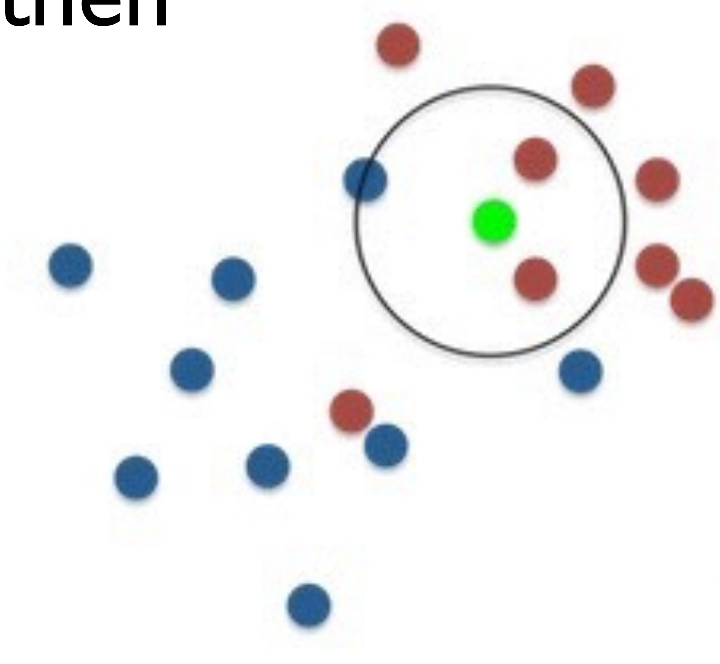
[Fix and Hodges'51, Stone'77, Cover and Hart 65,67,68]

Proof Intuition

Assume η is continuous

Theorem: If $k_n \to \infty$ and if $k_n/n \to 0$, then

$$R(h_{n,k_n}) \to R^* \quad \text{as} \quad n \to \infty$$

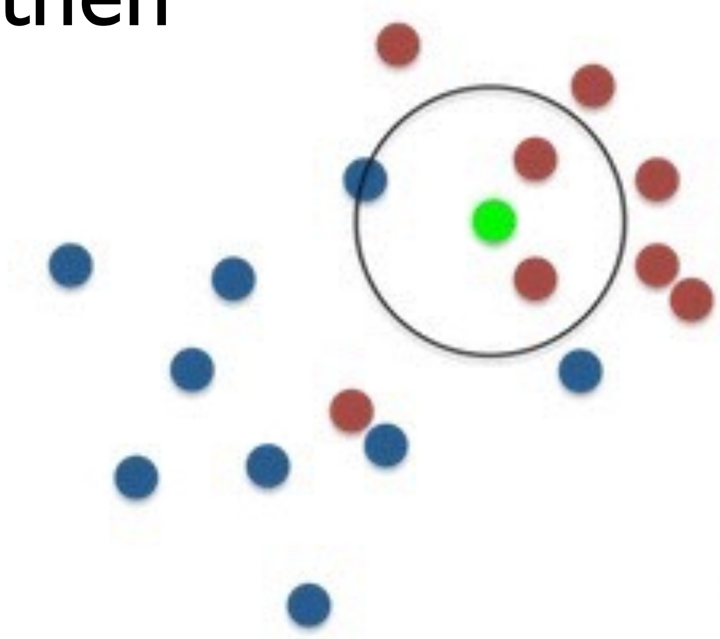


Proof Intuition

Assume η is continuous

Theorem: If $k_n \to \infty$ and if $k_n/n \to 0$, then

$$R(h_{n,k_n}) \to R^* \quad \text{as} \quad n \to \infty$$



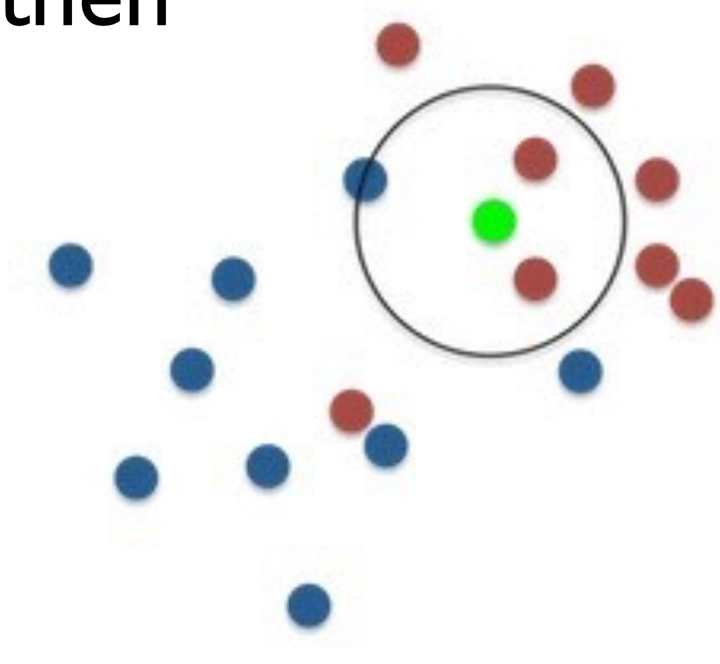
Proof: $X^{(1)}(x), \ldots, X^{(kn)}(x)$ lie in a ball of prob. mass $\approx k_n/n$

Proof Intuition

Assume η is continuous

Theorem: If $k_n \to \infty$ and if $k_n/n \to 0$, then

$$R(h_{n,k_n}) \to R^* \quad \text{as} \quad n \to \infty$$



Proof: $X^{(1)}(x), \ldots, X^{(kn)}(x)$ lie in a ball of prob. mass $\approx k_n/n$

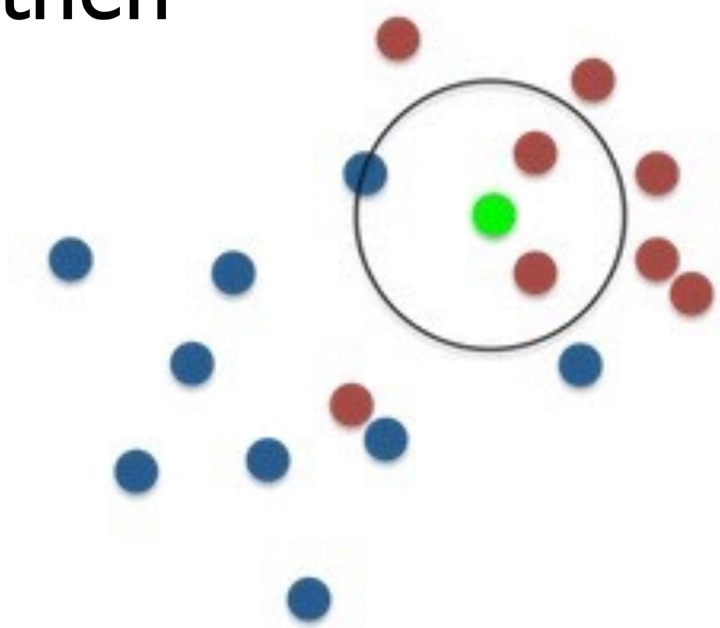
$X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$

Proof Intuition

Assume η is continuous

Theorem: If $k_n \to \infty$ and if $k_n/n \to 0$, then

$$R(h_{n,k_n}) \to R^* \quad \text{as} \quad n \to \infty$$



Proof: X(1)(x), …, X(kn)(x) lie in a ball of prob. mass $\approx k_n/n$

$X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$

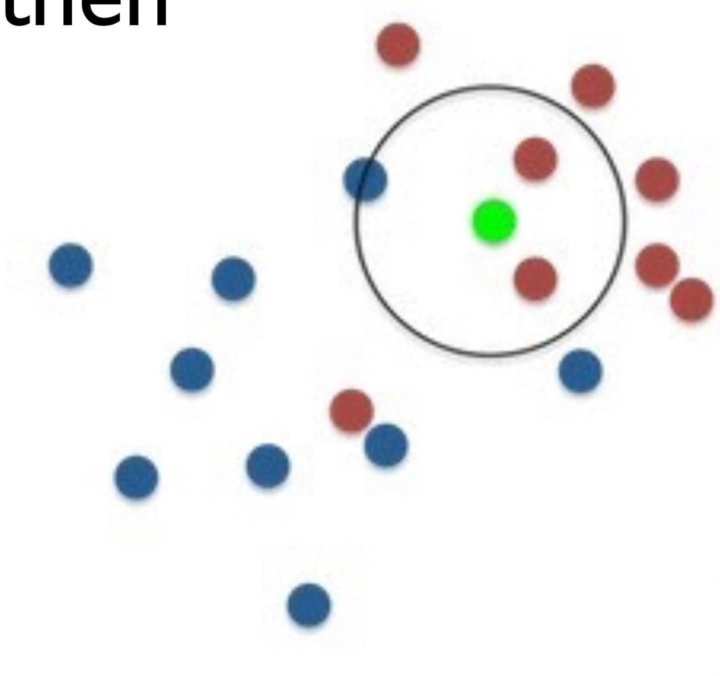
By continuity, $\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)) \to \eta(x)$

Proof Intuition

Assume η is continuous

Theorem: If $k_n \to \infty$ and if $k_n/n \to 0$, then

$$R(h_{n,k_n}) \to R^* \quad \text{as} \quad n \to \infty$$



Proof: $X^{(1)}(x), \ldots, X^{(kn)}(x)$ lie in

a ball of prob. mass $\approx k_n/n$

$X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$

By continuity, $\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)) \to \eta(x)$

As k_n grows, $\frac{1}{k_n}\sum_{i=1}^{k_n} Y^{(i)}(x) \to \eta(x)$

Universal Consistency in Metric Spaces

Theorem: Let (X, d, μ) be a separable metric measure space where the Lebesgue differentiation property holds:

For any bounded measurable f,

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

for almost all μ-a.e x in X

Universal Consistency in Metric Spaces

Theorem: Let (X, d, μ) be a separable metric measure space where the Lebesgue differentiation property holds:

For any bounded measurable f,

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

for almost all μ-a.e x in X

If $k_n \to \infty$ and $k_n/n \to 0$ then $R(h_{n,k_n}) \to R^*$ in probability

If in addition $k_n/\log n \to 0$ then $R(h_{n,k_n}) \to R^*$ almost surely

[Chaudhuri and Dasgupta, 14]

Universal Consistency in Metric Spaces

- Since $k_n/n \to 0$, $X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$

$\blacksquare$

Universal Consistency in Metric Spaces

- Since $k_n/n \to 0$, $X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$
- Earlier continuity argument: $\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)) \to \eta(x)$

$\blacksquare$

Universal Consistency in Metric Spaces

- Since $k_n/n \to 0$, $X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$
- Earlier continuity argument: $\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)) \to \eta(x)$
- It suffices that $\operatorname{avg}(\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x))) \to \eta(x)$

$\blacksquare$

Universal Consistency in Metric Spaces

- Since $k_n/n \to 0$, $X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$
- Earlier continuity argument: $\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)) \to \eta(x)$
- It suffices that $\mathrm{avg}(\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x))) \to \eta(x)$
- $X^{(1)}(x)$, .., $X^{(kn)}(x)$ lie in some ball B(x, r). For suitable r, they are random draws from μ restricted to B(x, r)

Universal Consistency in Metric Spaces

- Since $k_n/n \to 0$, $X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$
- Earlier continuity argument: $\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)) \to \eta(x)$
- It suffices that $\mathrm{avg}(\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x))) \to \eta(x)$
- $X^{(1)}(x), .., X^{(kn)}(x)$ lie in some ball B(x, r). For suitable r, they are random draws from μ restricted to B(x, r)
- $\mathrm{avg}(\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)))$ is close to avg η in B(x, r)

Universal Consistency in Metric Spaces

- Since $k_n/n \to 0$, $X^{(1)}(x), \ldots, X^{(k_n)}(x) \to x$
- Earlier continuity argument: $\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)) \to \eta(x)$
- It suffices that $\mathrm{avg}(\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x))) \to \eta(x)$
- $X^{(1)}(x), .., X^{(kn)}(x)$ lie in some ball B(x, r). For suitable r, they are random draws from μ restricted to B(x, r)
- $\mathrm{avg}(\eta(X^{(1)}(x)), \ldots, \eta(X^{(k_n)}(x)))$ is close to avg η in B(x, r)
- As n grows, this ball shrinks. Thus it is enough that

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

Tutorial Outline

- Nearest Neighbor Regression
    - The Setting
    - Universality
    - Rates of Convergence
- Nearest Neighbor Classification
    - The Statistical Learning Framework
    - Consistency
    - Rates of Convergence

Main Idea in Prior Analysis

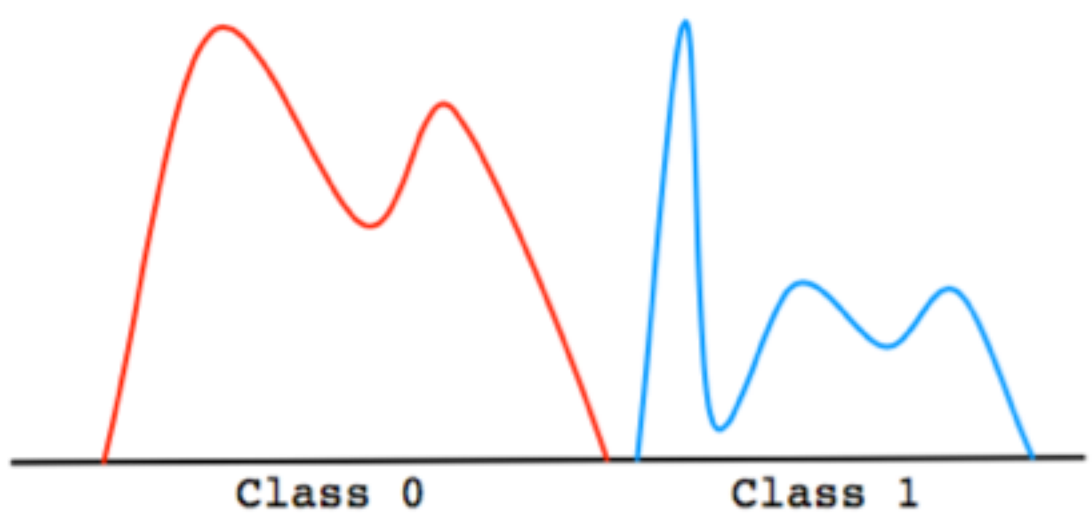
Smoothness of μ $\Rightarrow$ Small $r_k(x)$

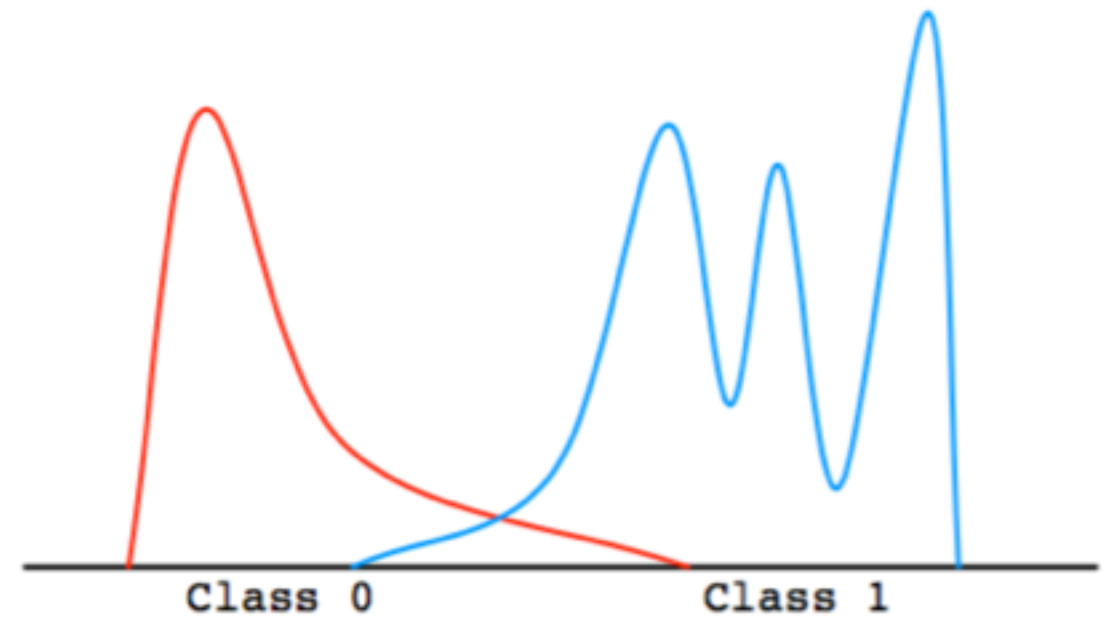
Lipschitzness of η $\Rightarrow$ $\eta(X^{(k)}(x)) \approx \eta(x)$

Neither smoothness nor Lipschitzness matter!

[Chaudhuri and Dasgupta'14]

A Motivating Example





Property of interest:

Balls of probability mass approx. k/n around x where x is close to the decision boundary

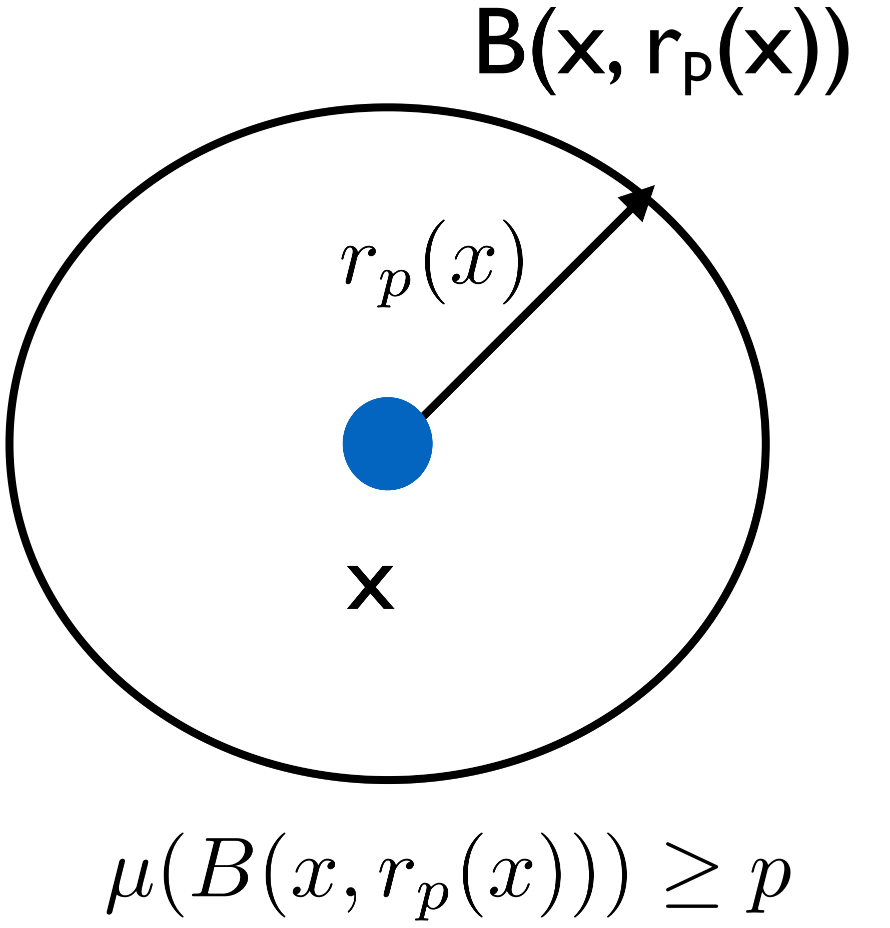
Some Notation

Probability-radius $r_p(x)$:

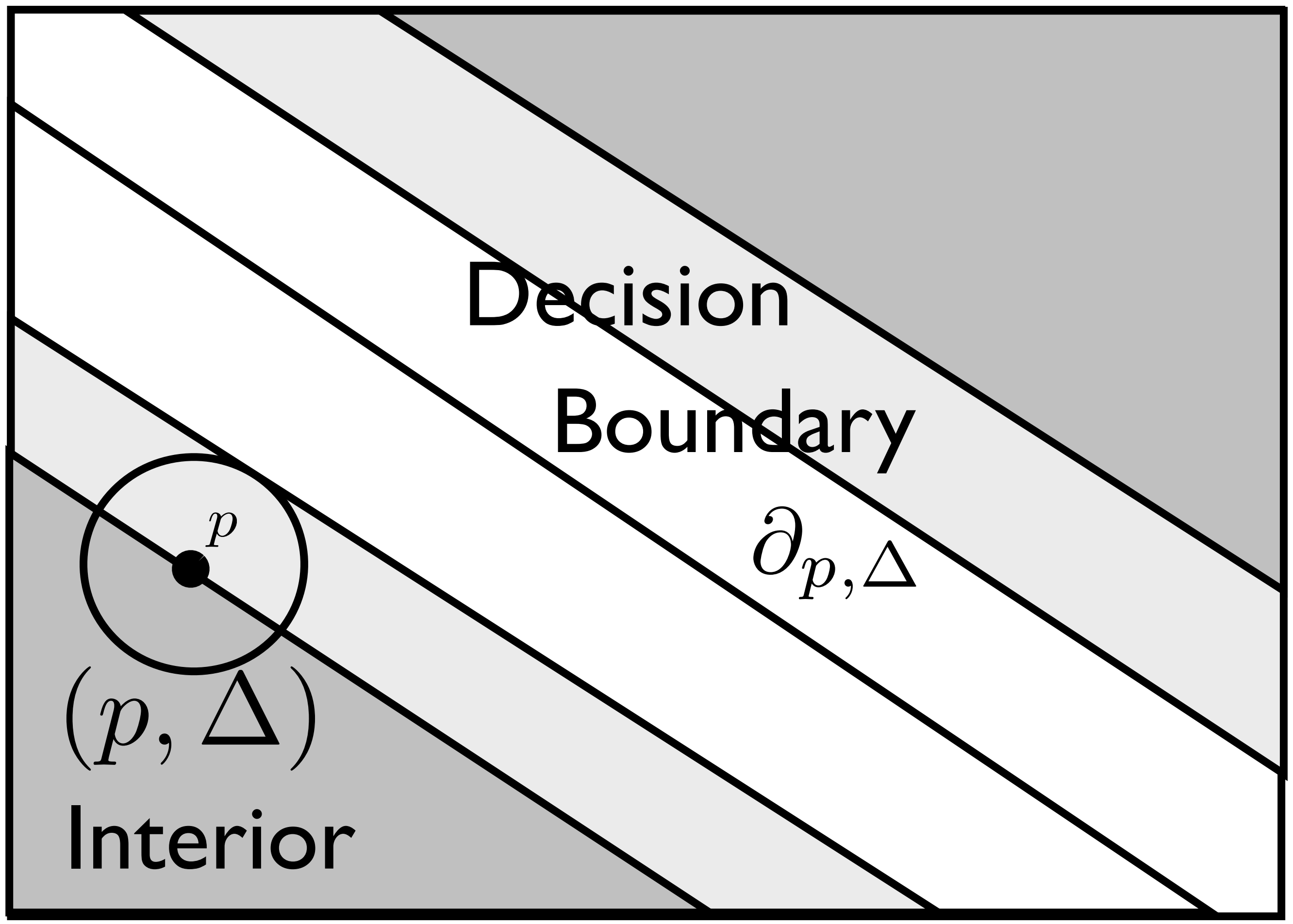
$$r_p(x) = \inf\{r | \mu(B(x,r)) \geq p\}$$

Conditional probability for a set:

$$\eta(A) = \frac{1}{\mu(A)} \int_A \eta d\mu$$



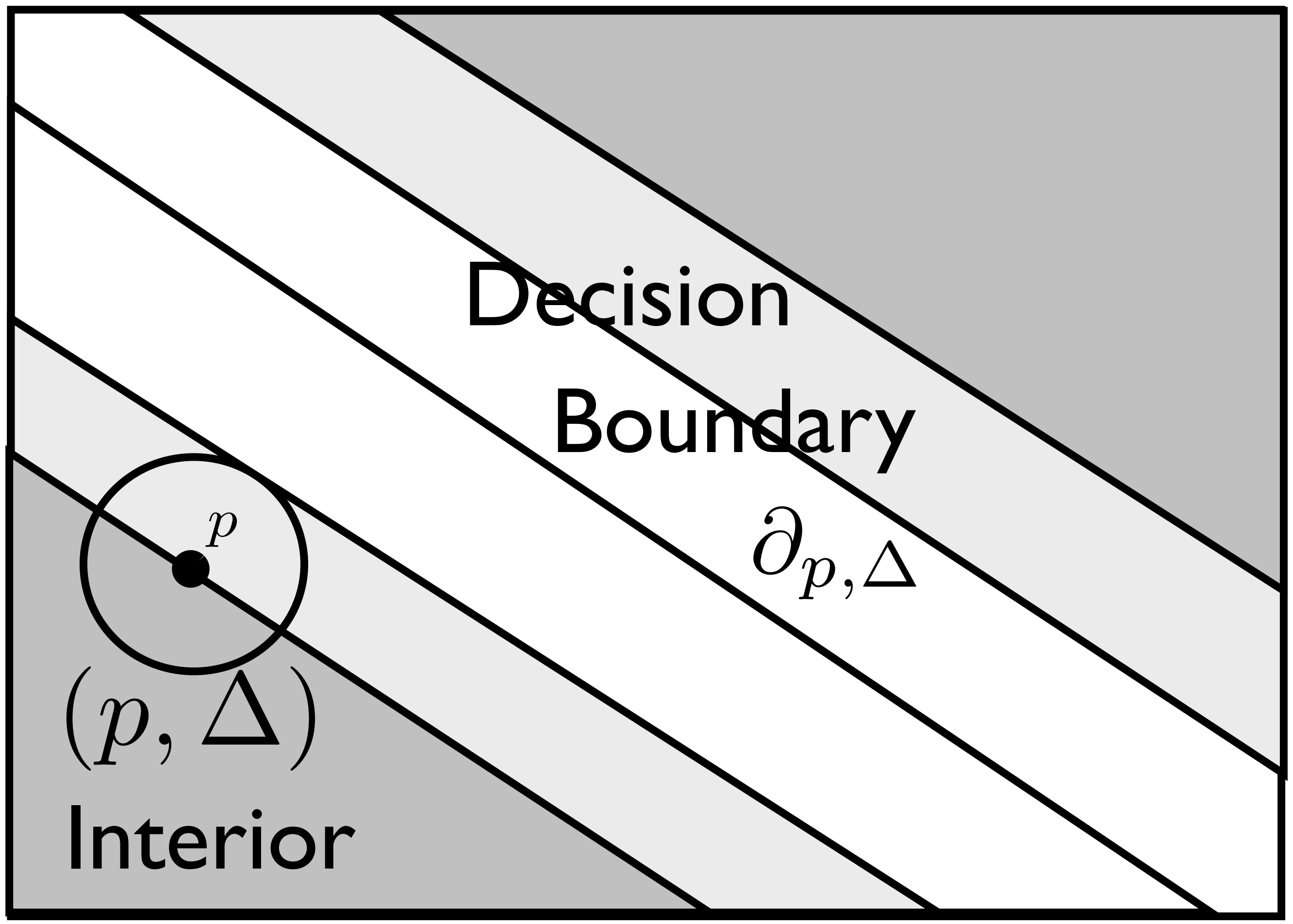
Effective Interiors and Boundaries



Positive Interior:

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

Effective Interiors and Boundaries

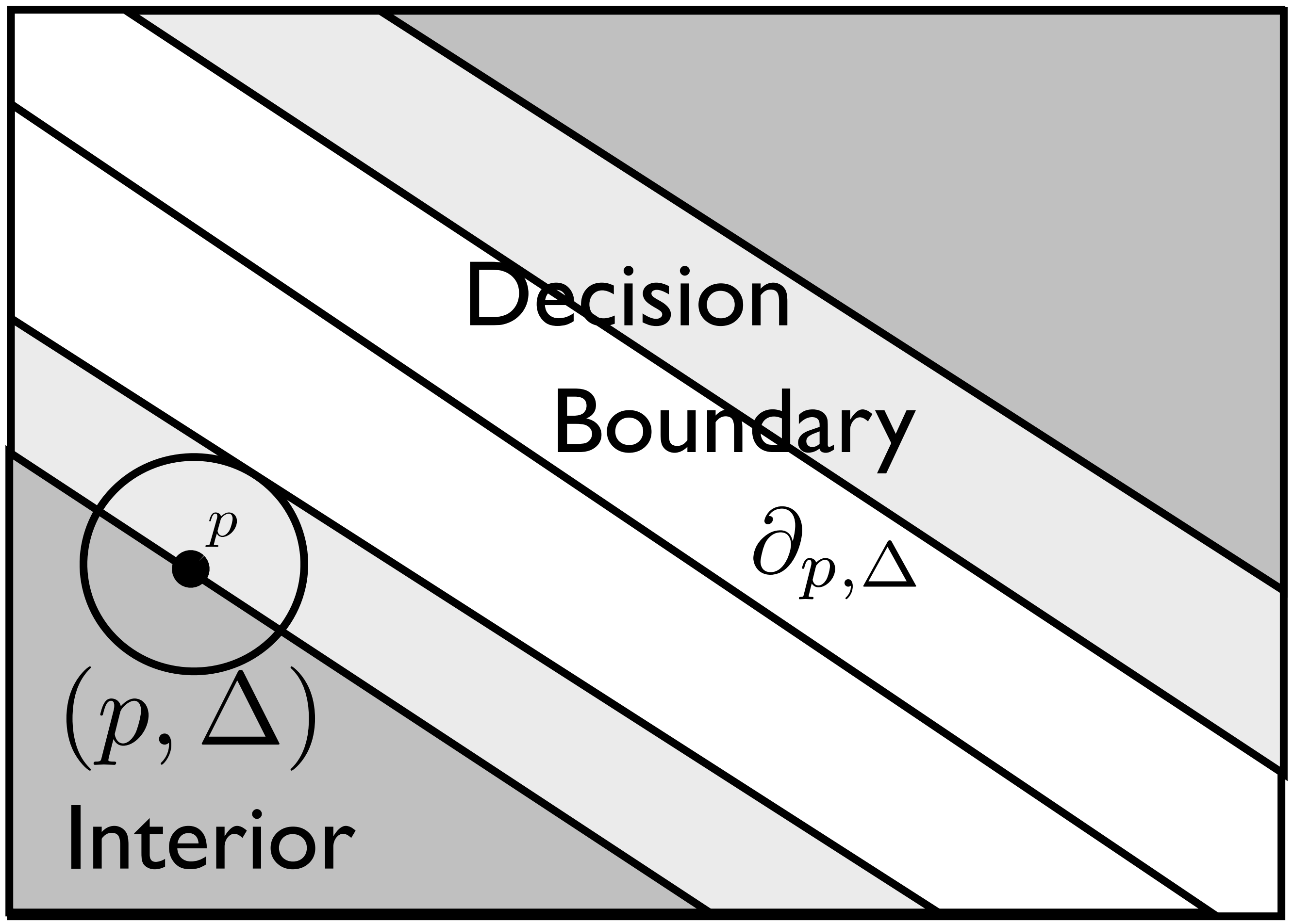


Positive Interior:

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

Similarly **Negative Interior**

Effective Interiors and Boundaries



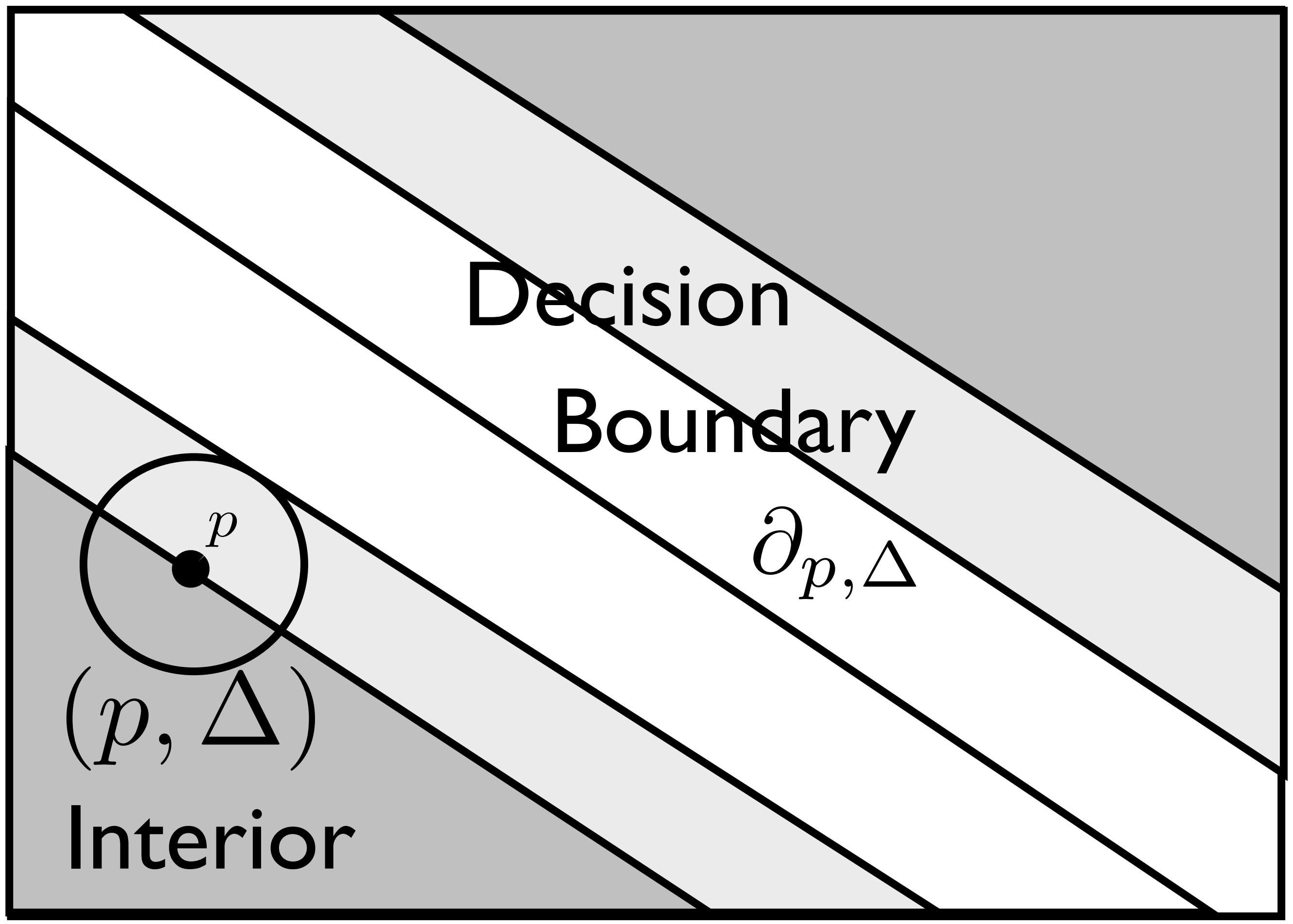
Positive Interior:

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

Similarly **Negative Interior**

(p,Δ) **-Interior:** $\mathcal{X}^+_{p,\Delta} \cup \mathcal{X}^-_{p,\Delta}$

Effective Interiors and Boundaries



Positive Interior:

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

Similarly **Negative Interior**

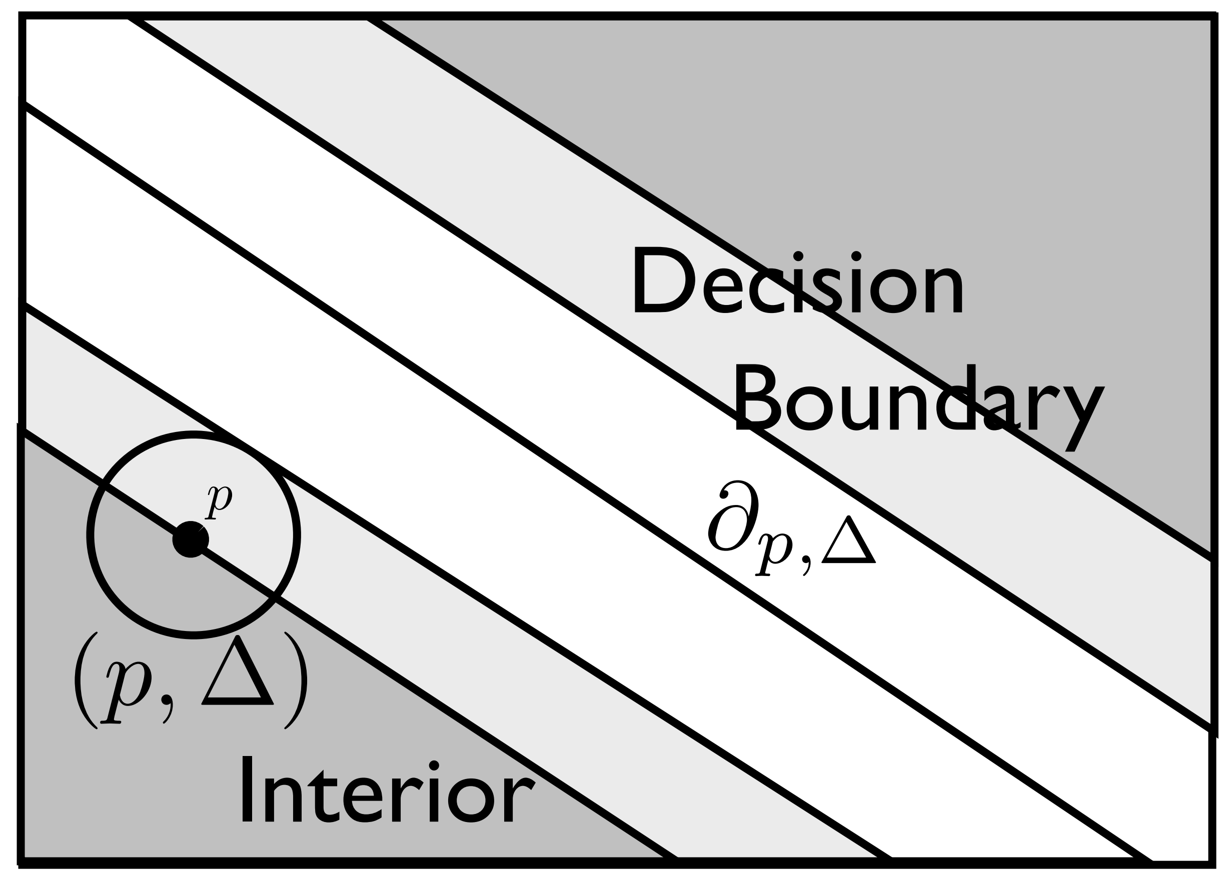
(p, Δ) **-Interior:** $\mathcal{X}^+_{p,\Delta} \cup \mathcal{X}^-_{p,\Delta}$

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

Convergence Rate Theorem

Risk $R_{n,k}$ of the k-NN classifier based on n training examples is:

$$R_{n,k} \leq R^* + \delta + \mu(\partial_{p,\Delta})$$



Convergence Rate Theorem

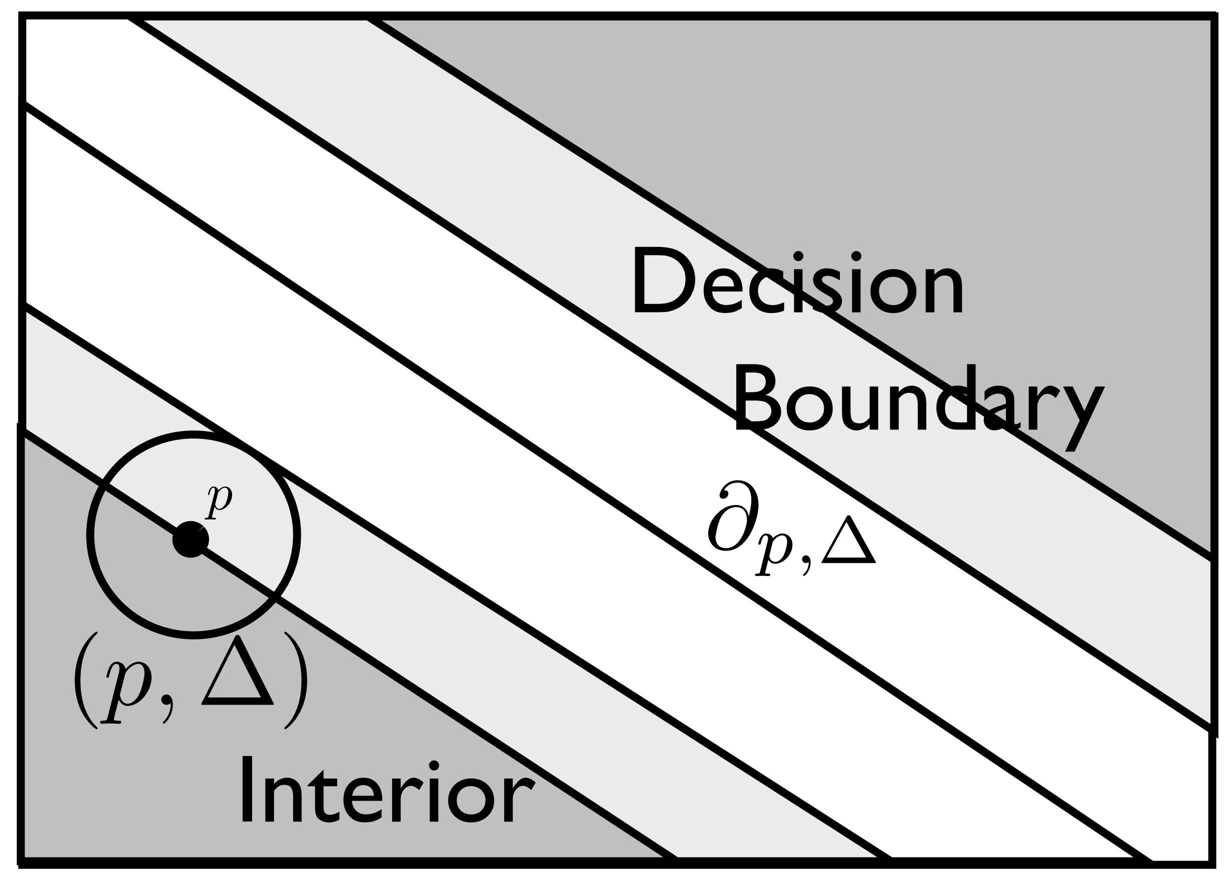
Risk $R_{n,k}$ of the k-NN classifier based on n training examples is:

$$R_{n,k} \leq R^* + \delta + \mu(\partial_{p,\Delta})$$

for:

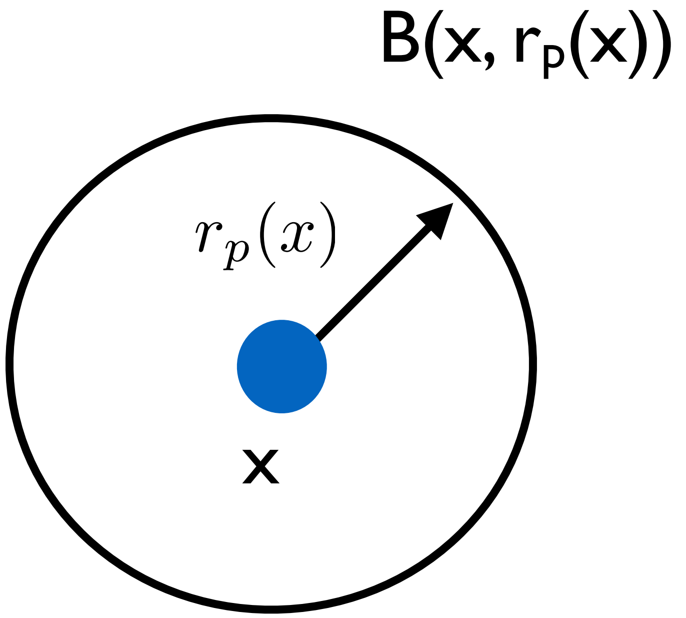
$$p = \frac{k}{n} \cdot \frac{1}{1 - \sqrt{(4/k)\log(2/\delta)}}$$

$$\Delta = \min\left(\frac{1}{2}, \sqrt{\frac{\log(2/\delta)}{k}}\right)$$



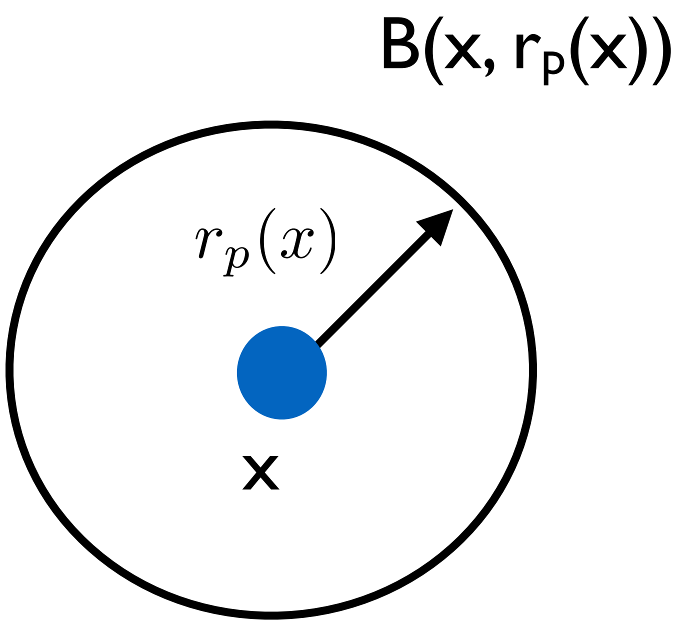
Proof Intuition 1

For fixed x, let B = B(x, r_p(x))



If $h_{n,k}(x) \neq h(x)$ then:

Proof Intuition I

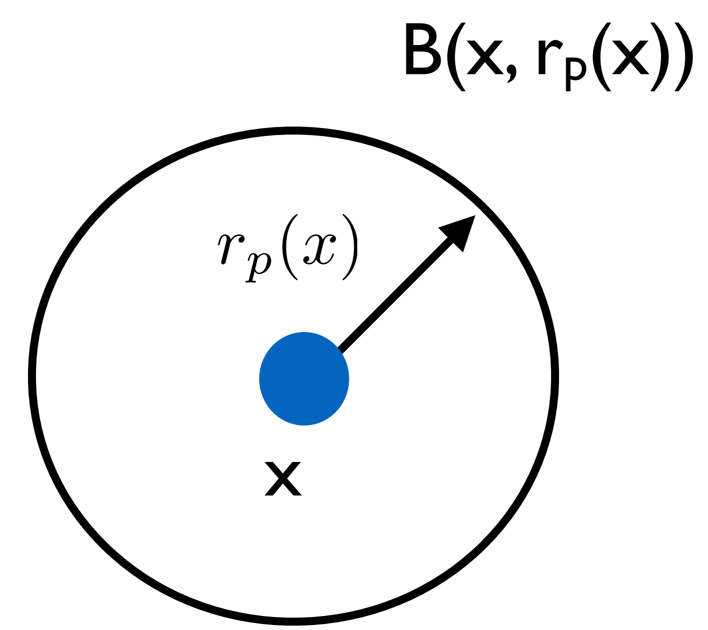


For fixed x, let B = B(x, r_p(x))

If $h_{n,k}(x) \neq h(x)$ then:

1. $x \in \partial_{p,\Delta}$

Proof Intuition I

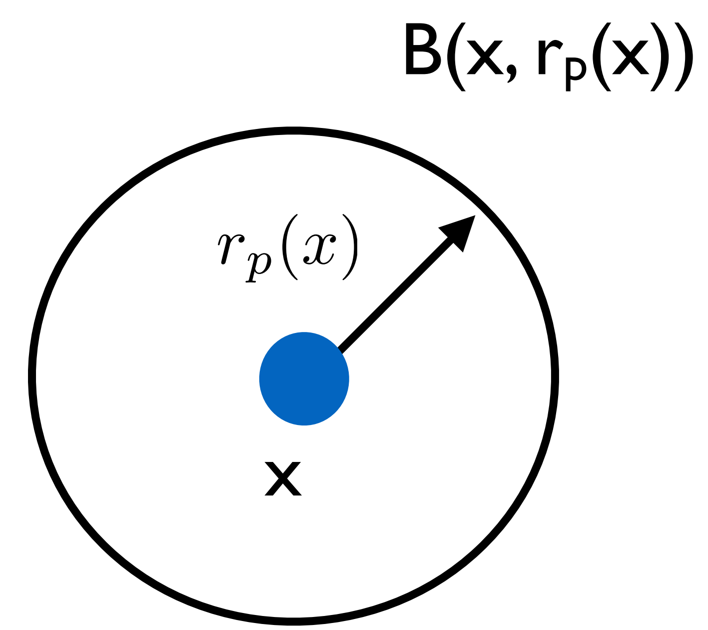


For fixed x, let B = $B(x, r_p(x))$

If $h_{n,k}(x) \neq h(x)$ then:

1. $x \in \partial_{p,\Delta}$
2. $d(x, X^{(k)}(x)) > r_p(x)$

Proof Intuition I

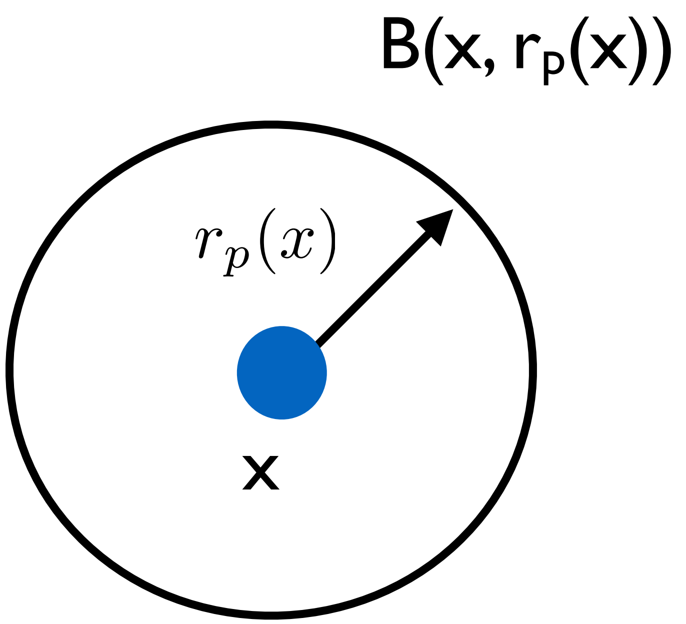


For fixed x, let B = B(x, r_p(x))

If $h_{n,k}(x) \neq h(x)$ then:

1. $x \in \partial_{p,\Delta}$
2. $d(x, X^{(k)}(x)) > r_p(x)$
3. $\left| \frac{1}{|B|} \sum_i Y_i \cdot 1(X_i \in B) - \eta(B) \right| \geq \Delta$

Proof Intuition I



For fixed x, let B = B(x, r_p(x))

If $h_{n,k}(x) \neq h(x)$ then:

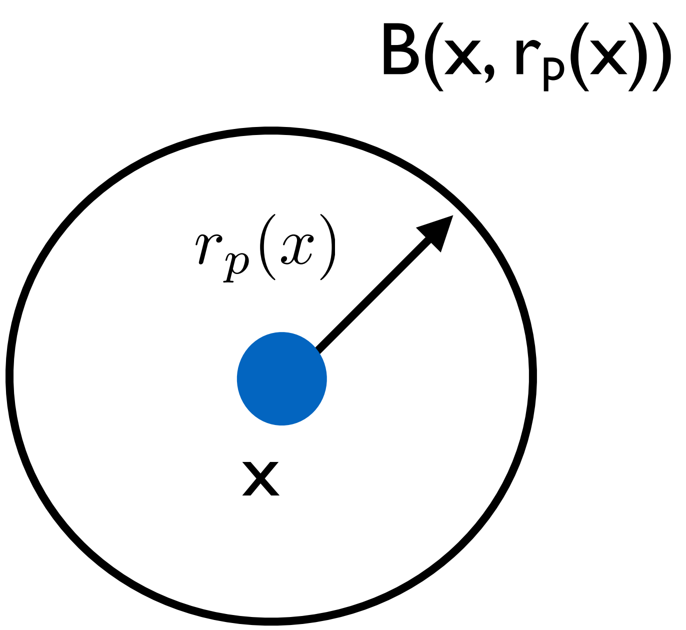
1. $x \in \partial_{p,\Delta}$
2. $d(x, X^{(k)}(x)) > r_p(x)$
3. $\left|\frac{1}{|B|}\sum_i Y_i \cdot 1(X_i \in B) - \eta(B)\right| \geq \Delta$

If (1) does not hold, say

$\eta(x) \geq 1/2$

Then $\eta(B) \geq 1/2 + \Delta$

Proof Intuition 1



For fixed x, let B = B(x, r_p(x))

If $h_{n,k}(x) \neq h(x)$ then:

1. $x \in \partial_{p,\Delta}$
2. $d(x, X^{(k)}(x)) > r_p(x)$
3. $\left| \frac{1}{|B|} \sum_i Y_i \cdot 1(X_i \in B) - \eta(B) \right| \geq \Delta$

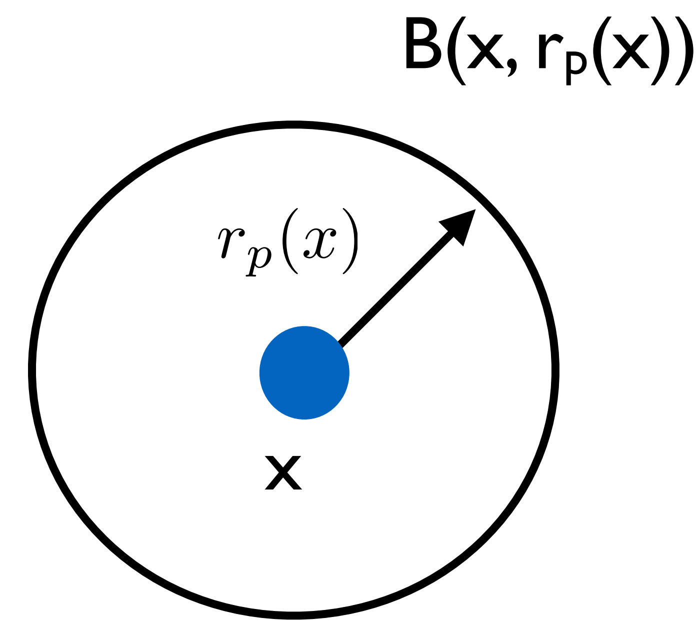
If (1) does not hold, say

$\eta(x) \geq 1/2$

Then $\eta(B) \geq 1/2 + \Delta$

Either k-th NN of x lies outside B or (3) holds

Proof Intuition 2



For fixed x, let B = B(x, r_p(x))

If $h_{n,k}(x) \neq h(x)$ then:

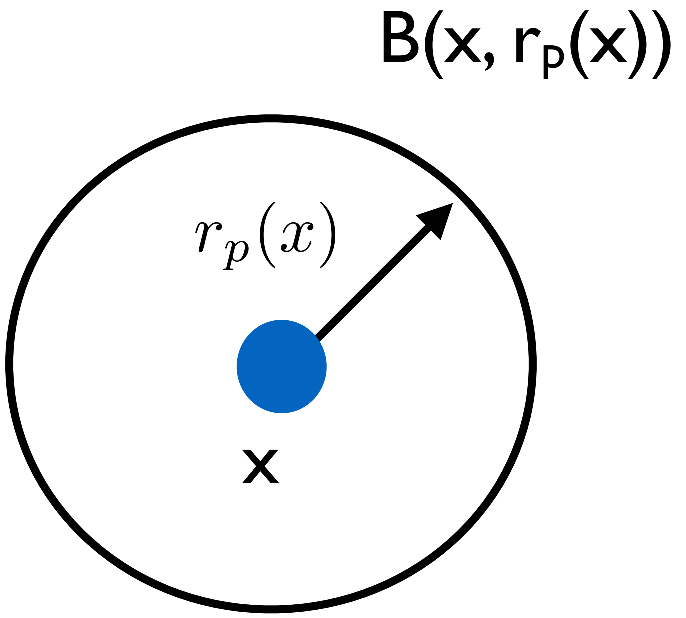
1. $x \in \partial_{p,\Delta}$
2. $d(x, X^{(k)}(x)) > r_p(x)$
3. $\left| \frac{1}{|B|} \sum_i Y_i \cdot 1(X_i \in B) - \eta(B) \right| \geq \Delta$

If

$$p = \frac{k}{n} \cdot \frac{1}{1 - \sqrt{(4/k)\log(2/\delta)}}$$

then, the probability of (2) is at most $\delta/2$ (Chernoff bounds)

Proof Intuition 3



For fixed x, let B = B(x, r_p(x))

If $h_{n,k}(x) \neq h(x)$ then:

1. $x \in \partial_{p,\Delta}$
2. $d(x, X^{(k)}(x)) > r_p(x)$
3. $\left| \frac{1}{|B|} \sum_i Y_i \cdot 1(X_i \in B) - \eta(B) \right| \geq \Delta$

If

$$\Delta = \min\left(\frac{1}{2}, \sqrt{\frac{\log(2/\delta)}{k}}\right)$$

then, the probability of (3) is at most $\delta/2$ (Chernoff bounds)

Putting it all together…

Risk $R_{n,k}$ of the k-NN classifier based on n training examples is:

$$R_{n,k} \leq \Pr(h(x) \neq y) + \Pr(x \in \partial_{p,\Delta}) + \Pr(2.) + \Pr(3.)$$

Putting it all together…

Risk $R_{n,k}$ of the k-NN classifier based on n training examples is:

$$R_{n,k} \leq \Pr(h(x) \neq y) + \Pr(x \in \partial_{p,\Delta}) + \Pr(2.) + \Pr(3.)$$

$$\Pr(h(x) \neq y) = R^*$$ (By definition)

Putting it all together…

Risk $R_{n,k}$ of the k-NN classifier based on n training examples is:

$$R_{n,k} \leq \Pr(h(x) \neq y) + \Pr(x \in \partial_{p,\Delta}) + \Pr(2.) + \Pr(3.)$$

$$\Pr(h(x) \neq y) = R^*$$ (By definition)

$$\Pr(x \in \partial_{p,\Delta}) = \mu(\partial_{p,\Delta})$$

Putting it all together…

Risk $R_{n,k}$ of the k-NN classifier based on n training examples is:

$$R_{n,k} \leq \Pr(h(x) \neq y) + \Pr(x \in \partial_{p,\Delta}) + \Pr(2.) + \Pr(3.)$$

$$\Pr(h(x) \neq y) = R^*$$ (By definition)

$$\Pr(x \in \partial_{p,\Delta}) = \mu(\partial_{p,\Delta})$$

If $p = \frac{k}{n} \cdot \frac{1}{1 - \sqrt{(4/k)\log(2/\delta)}}$ and $\Delta = \min\left(\frac{1}{2}, \sqrt{\frac{\log(2/\delta)}{k}}\right)$

then $\Pr(2.) + \Pr(3.) \leq \delta$

Convergence Rate Theorem

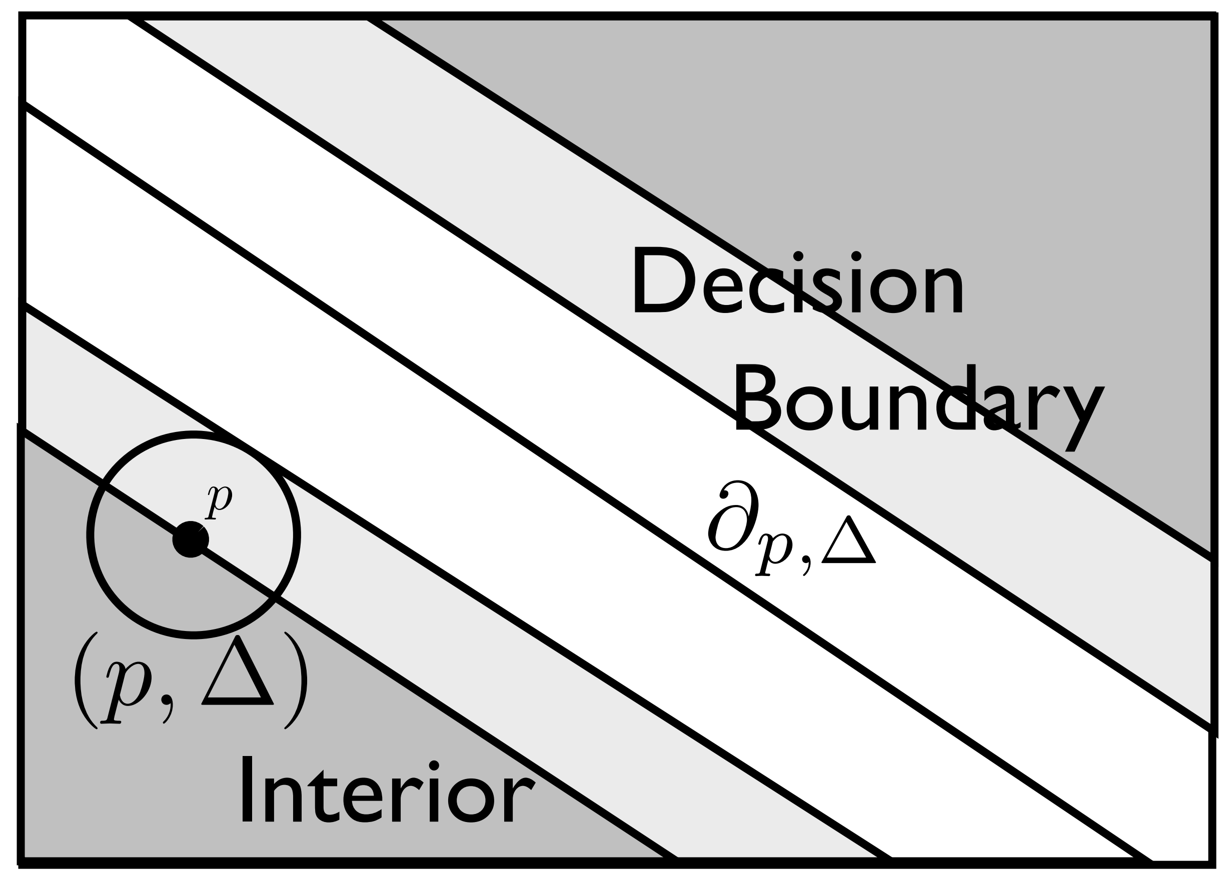
Risk $R_{n,k}$ of the k-NN classifier based on n training examples is:

$$R_{n,k} \leq R^* + \delta + \mu(\partial_{p,\Delta})$$

for:

$$p = \frac{k}{n} \cdot \frac{1}{1 - \sqrt{(4/k)\log(2/\delta)}}$$

$$\Delta = \min\left(\frac{1}{2}, \sqrt{\frac{\log(2/\delta)}{k}}\right)$$



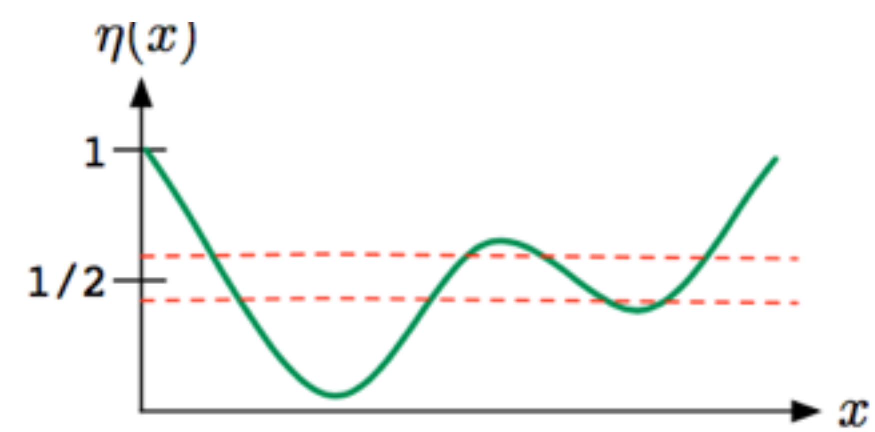
Smoothness

η is α-Holder continuous if for constant L, all x, x',

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

Margin: For constant C, for any t,

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



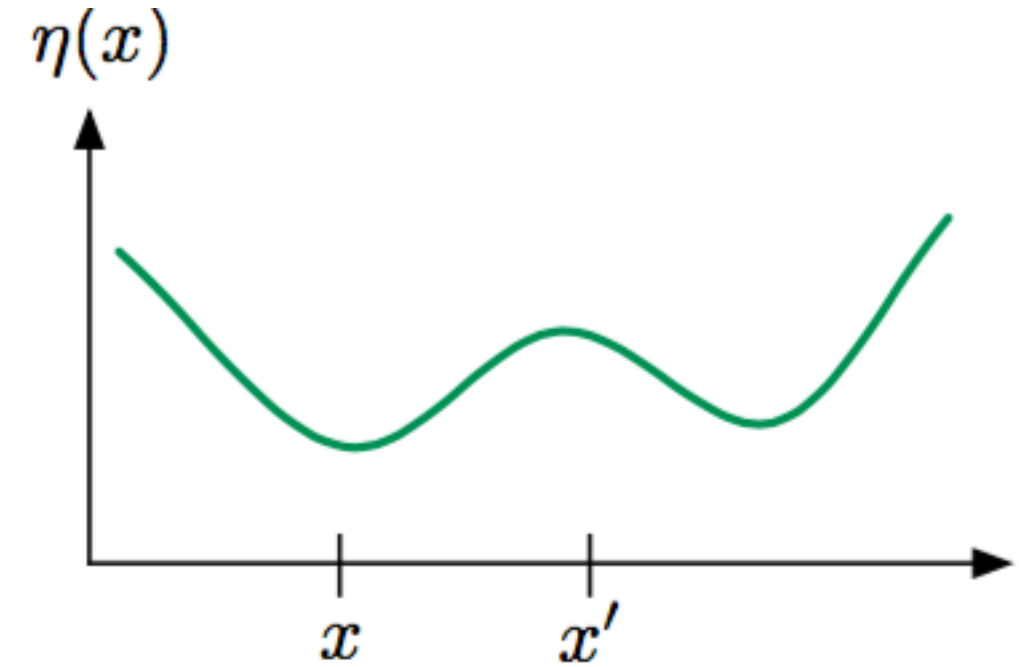
The above two conditions plus μ is supported on a regular set with $\mu_{\min} \leq \mu \leq \mu_{\max}$

Then E[R] - R* is $\Theta(n^{-\alpha(\beta+1)/(2\alpha+d)})$

Also achieved by k-NN for suitable k

A Better Smoothness Condition

More natural notion:
Relate smoothness to
$\mu(\|x - x'\|)$



η is α-smooth if for some constant L, for all x, r > 0,

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

Smoothness Bounds

Suppose η is α -smooth. Then for any n, k,

With probability $\geq 1-\delta$,

$$\Pr(h_{n,k}(X) \neq h(X)) \leq \delta + \mu\left(\{x| \ |\eta(x) - 1/2| \leq C_1\sqrt{\frac{1}{k}\log\frac{1}{\delta}}\right)$$

For $k \propto n^{2\alpha/(2\alpha+1)}$

Lower Bounds: With constant probability,

$$\Pr(h_{n,k}(X) \neq h(X)) \geq C_2\mu\left(\{x| \ |\eta(x) - 1/2| \leq C_3\sqrt{\frac{1}{k}}\}\right)$$

Implications

1. Recovers previous bounds on smooth functions with margin conditions

2. Faster rates for special cases

    - Zero Bayes Risk: 1-NN has the best rates

    - Δ Bounded away from 0: Exponential convergence

Conclusion

1. k_n-NN is always universally consistent provided k grows a certain way with n

2. k-NN regression suffers from curse of dimensionality

3. k-NN classification also does, but can do better

Acknowledgements

Thanks to Sanjoy Dasgupta and Samory Kpotufe
A chunk of this talk is based on a tutorial
from ICML 2018 by Sanjoy and Samory