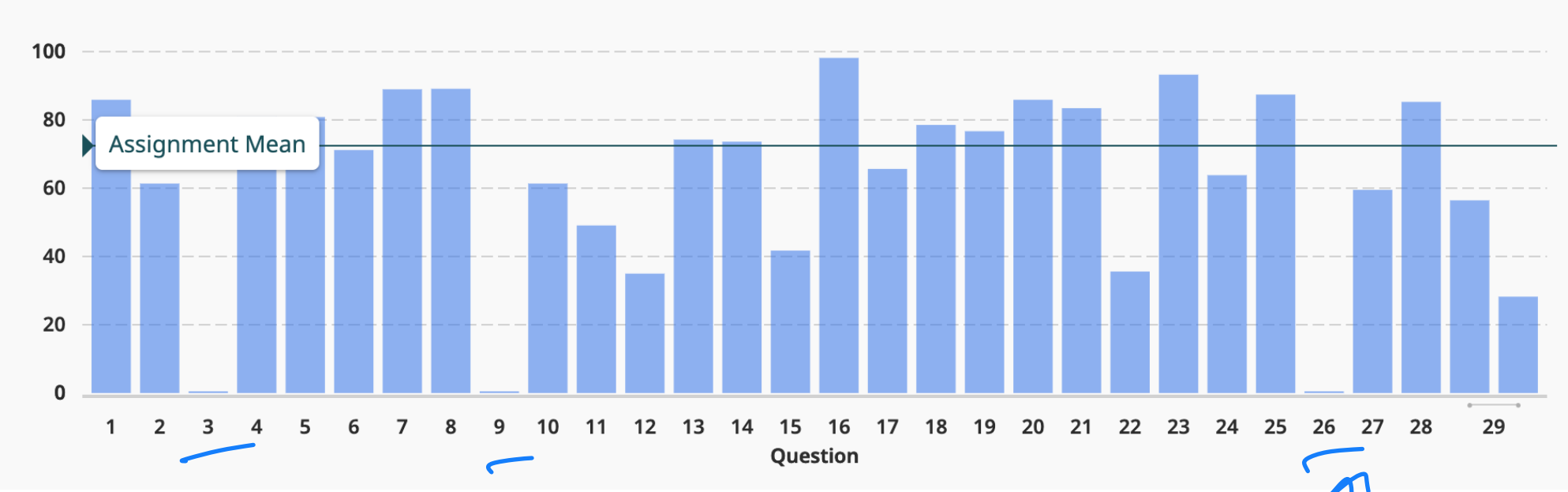
Midterm grades adjusted

- 3 hardest questions became bonus questions



Midterm Exam 27.0 points

Minimum	Median	Maximum	Mean	Std Dev
34.26%	73.15%	108.33%	72.22%	16.14%

Non-parametric methods
Nearest Neighbours

Natasha Jaques

UNIVERSITY of WASHINGTON

Parametric vs non-parametric

- A model is parametric if # parameters does not depend on # samples
- A model is non-parametric if # parameters increases with # samples
  - Does not mean absence of parameters!

This lecture: k nearest neighbors

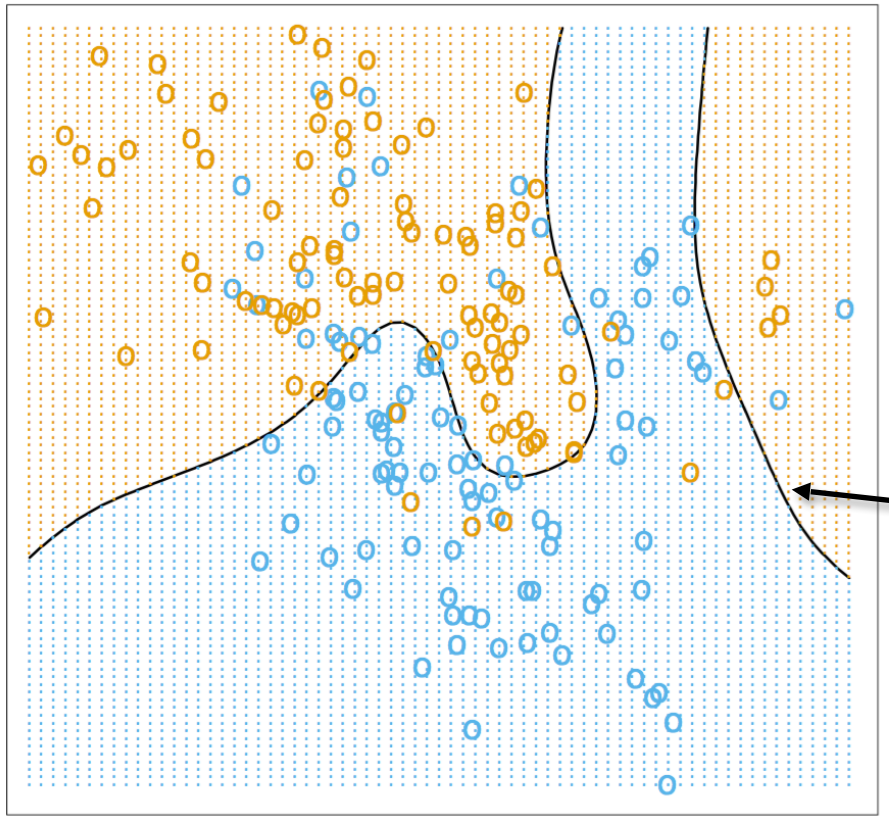
simple

- Assume we have a classification task
- To classify a new point x:
  - Find its k nearest neighbors in the training data
  - Set y to be the majority vote of the labels of these nearest neighbors

smallest distance in feature space (2 features)

- Design choices / hyperparameters:
  - Number of nearest neighbors k
  - Distance metric
  - Aggregation method

Example: Bayes classifier



Training data:

- True label: +1
- True label: -1

Optimal Bayes classifier:

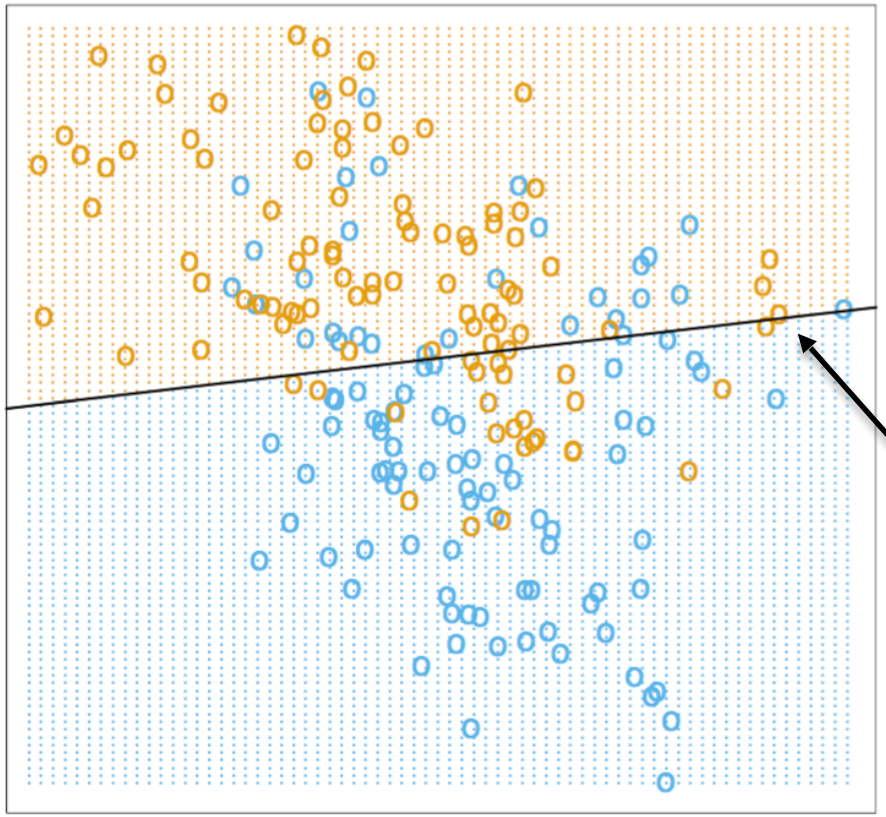
$$\mathbb{P}(Y = 1|X = x) = \frac{1}{2}$$

← need function

- Predicted label: +1
- Predicted label: -1

Figures from Hastie et al.

Linear decision boundary



Training data:

- True label: +1
- True label: -1

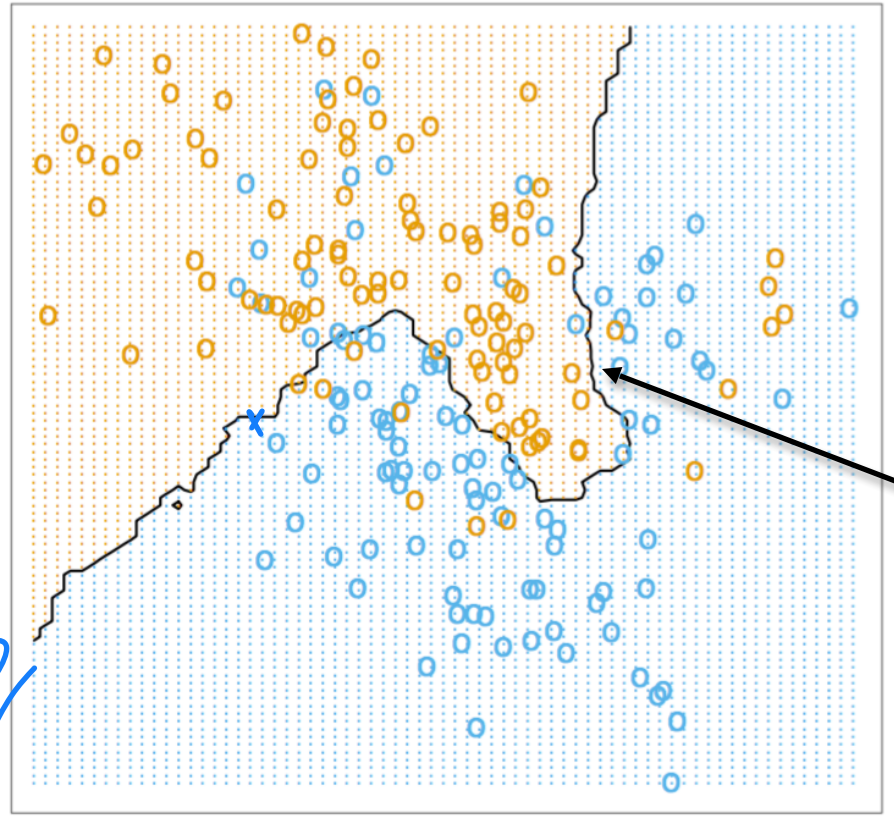
Learned linear decision boundary:

$$x^T w + b = 0$$

- Predicted label: +1
- Predicted label: -1

Figures from Hastie et al.

$k = 15$ nearest neighbors boundary



Training data:

True label: +1

True label: -1

15 nearest neighbors decision boundary (majority vote)

Predicted label: +1

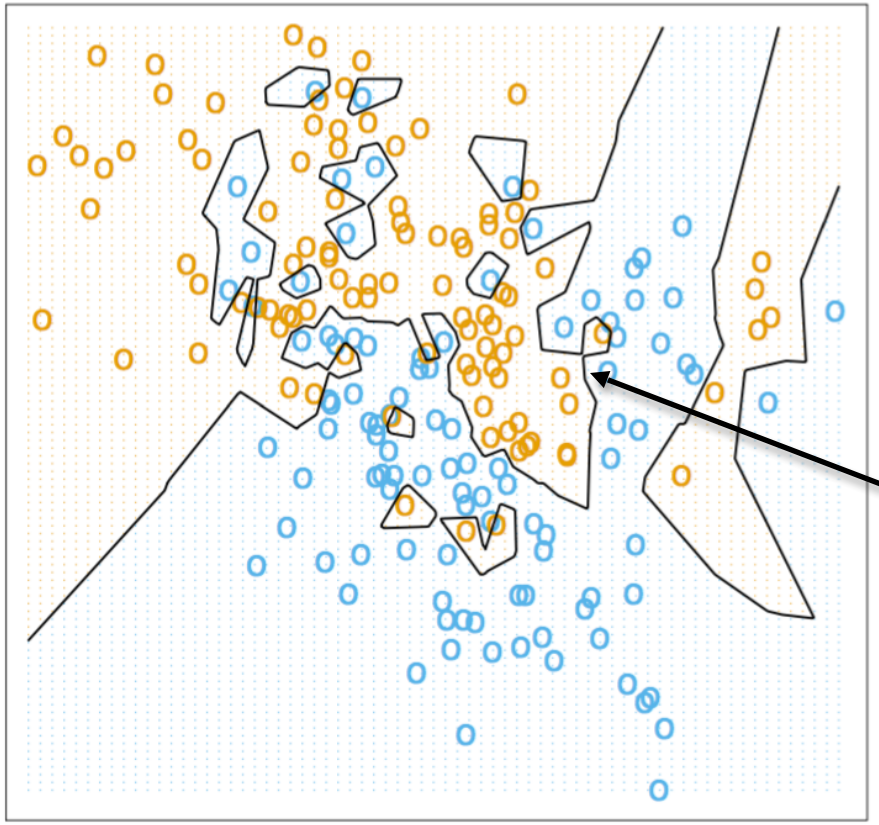
Predicted label: -1

boundary depends on the training data points

Figures from Hastie et al.

$k = 1$ nearest neighbor boundary

overfitting



Training data:

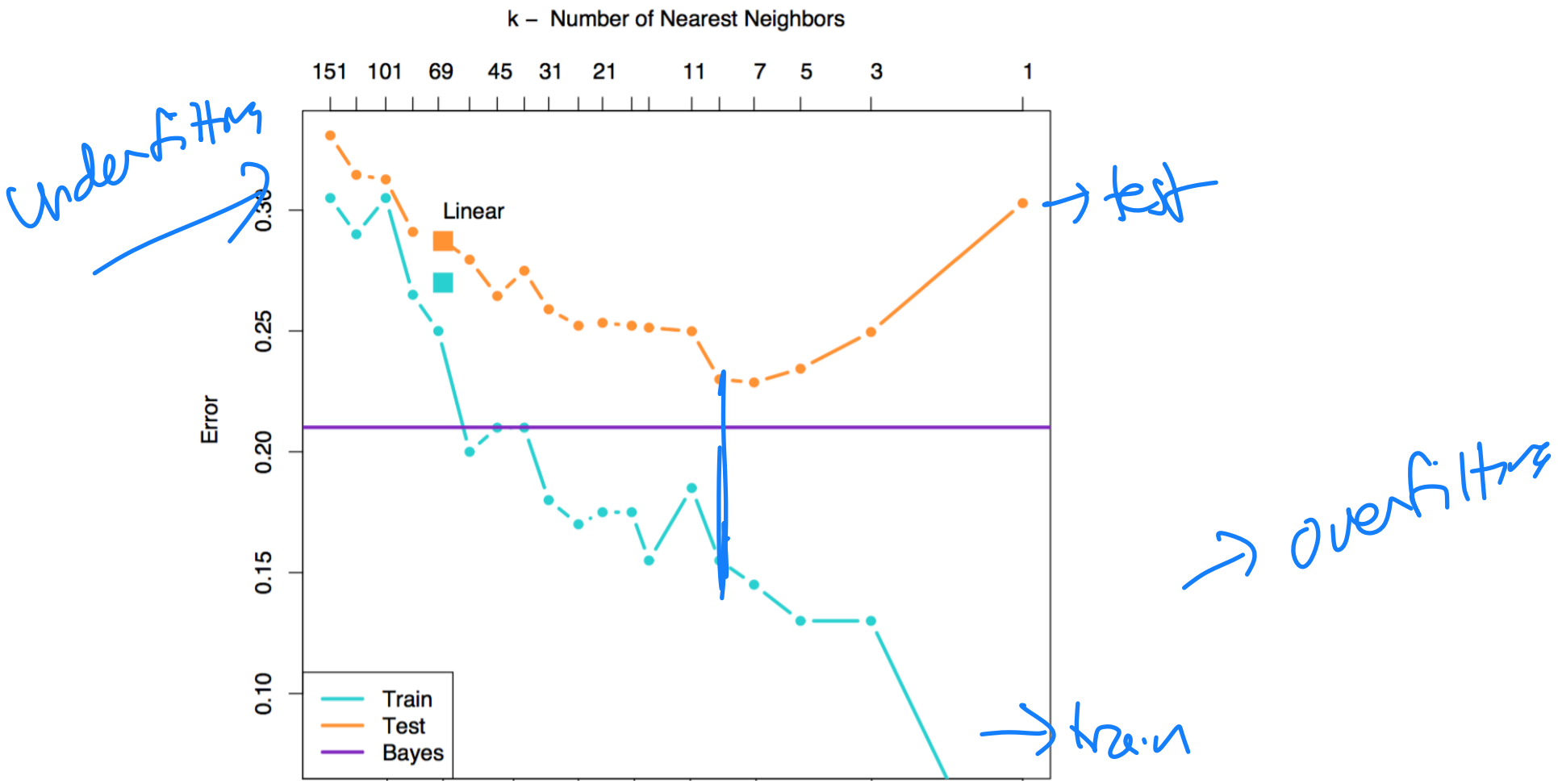
- True label: +1
- True label: -1

1 nearest neighbor decision boundary (majority vote)

- Predicted label: +1
- Predicted label: -1

Figures from Hastie et al.

k nearest neighbors error



Figures from Hastie et al.

Parametric vs non-parametric

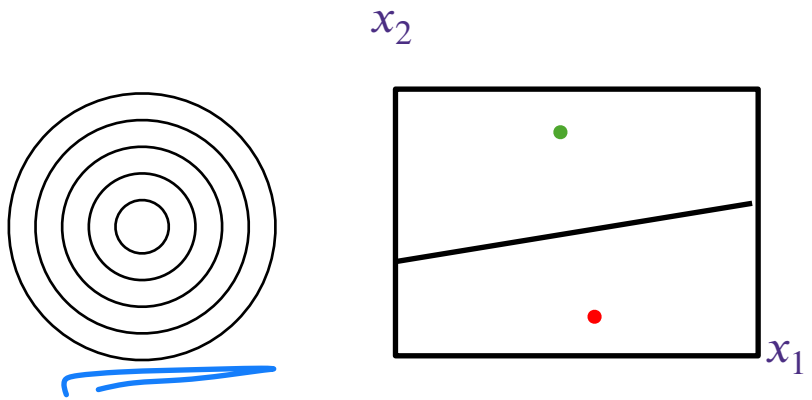
After training, discard your data

keep training data around for inference
eg. kNN

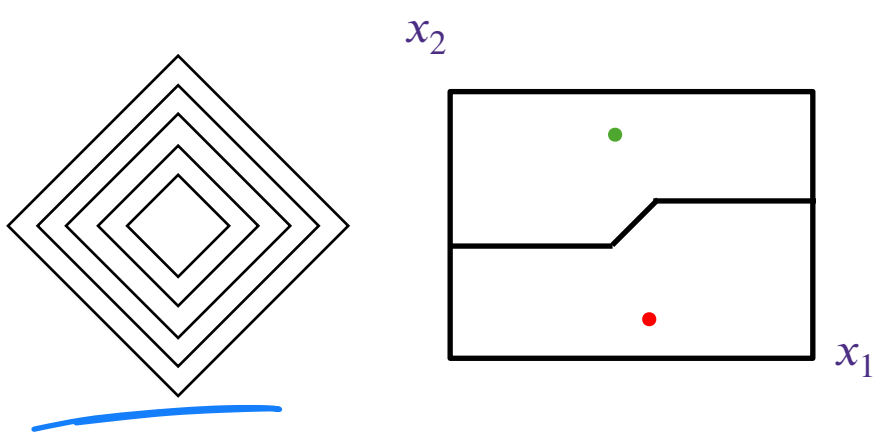
interpretable

Notable distance metrics & level sets

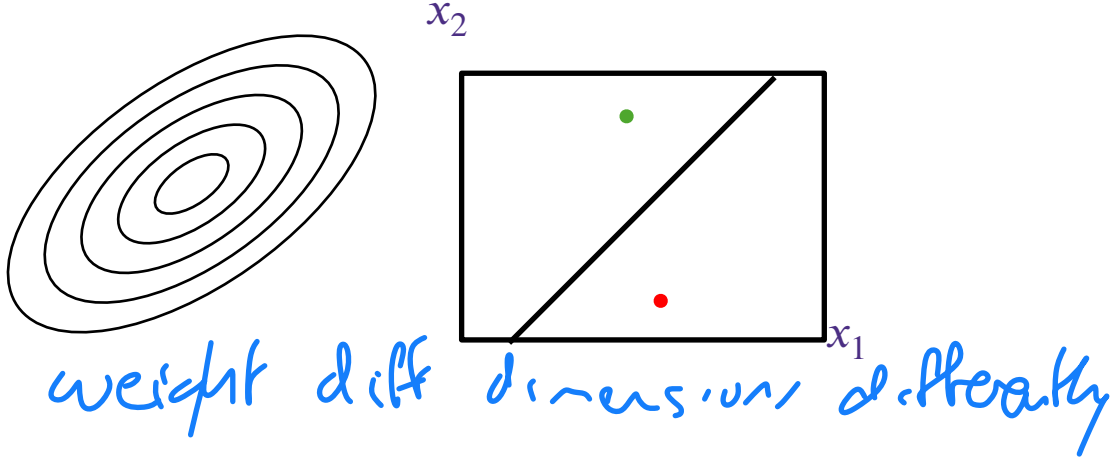
ℓ_2 norm (Euclidean)



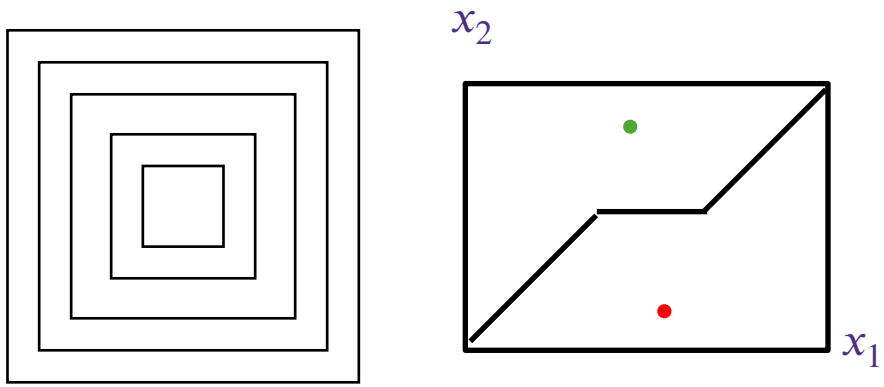
ℓ_1 norm (Manhattan, taxicab)



Mahalanobis norm



ℓ_∞ norm (max)

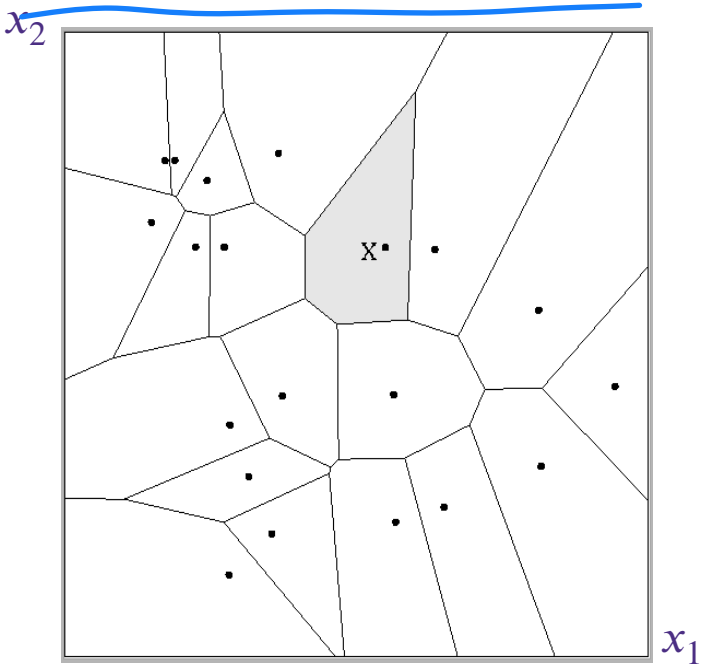


Example: distance metrics with $k = 1$ NN

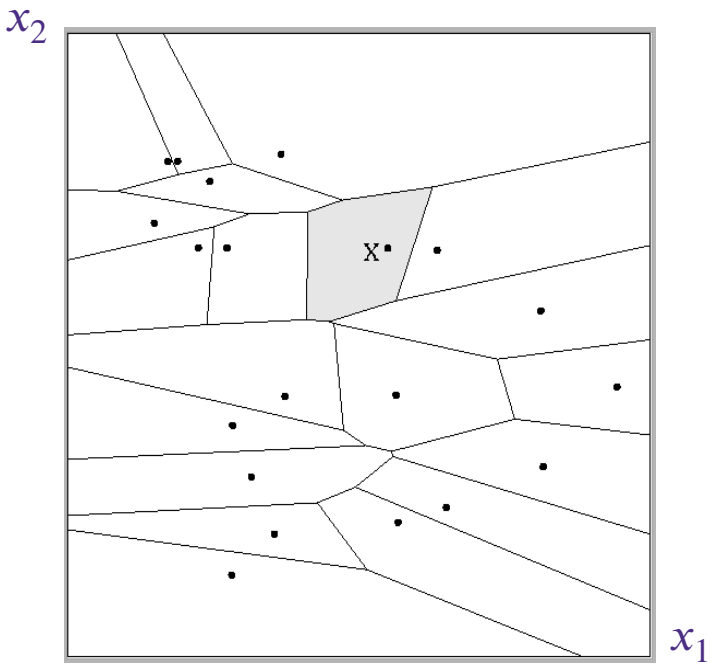
l2

Mahalanobis

$$d(x, x') = (x_1 - x'_1)^2 + (x_2 - x'_2)^2$$



$$d(x, x') = (x_1 - x'_1)^2 + 9(x_2 - x'_2)^2$$



//know geometry of feat space

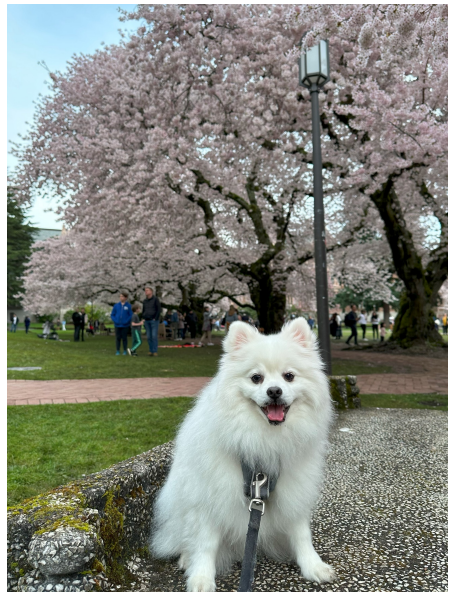
Learned distance metrics

Training data

Dog

Cat

Test data



learn distance with nn

Cool online kNN demo (credit to Vicky Ye)

http://vision.stanford.edu/teaching/cs231n-demos/knn/

1-NN classification: Theoretical guarantees

have enough data → function is smooth

$D = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n} \sim P \quad x^{(i)} \in \mathbb{R}^d, \; y \in \{0, 1\}$

Given test point x, let x_{NN} be the nearest neighbour in D

Error if $y_{NN} \neq y$

Case 1: $y_{NN} = 1, \; y = 0$ w.p. $p(y=1 \mid x_{NN})\, p(y=0 \mid x)$

Case 2: $y_{NN} = 0, \; y = 1$ w.p. $p(y=0 \mid x_{NN})\, p(y=1 \mid x)$

As $n \to \infty$, $p(y \mid x_{NN}) \to p(y \mid x)$ // distance b/w x & $x_{NN} \to 0$ as $n \to \infty$

Error: $2\, p(y=1 \mid x)\, p(y=0 \mid x)$ as $n \to \infty$

$= 2\, p^*(1 - p^*)$

$\leq 2\, p^*$

Bayes error

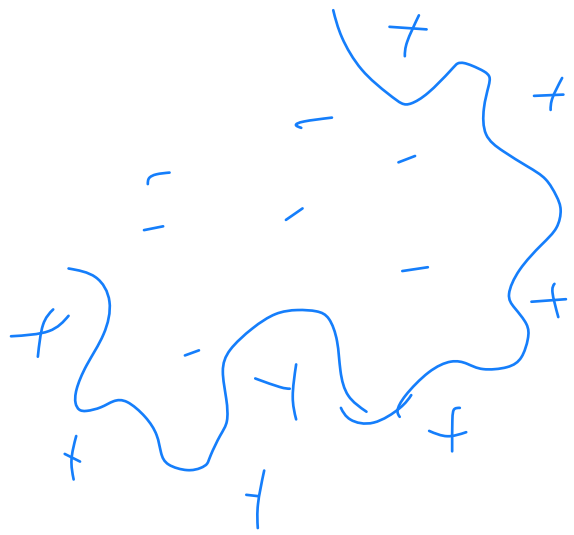
define $p^* = \min\{p(y=1 \mid x), p(y=0 \mid x)\}$

best error I can get ≠ measurement error

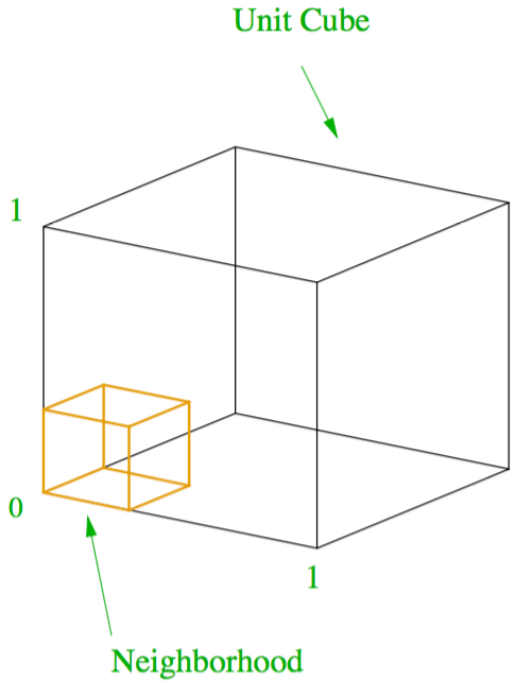
1-NN classification: Theoretical guarantees

Theorem[Cover, Hart, 1967] If P_X is supported everywhere in $\mathbb{R}^d$ and $P(Y = 1|X = x)$ is smooth everywhere, then as $n \to \infty$ the 1-NN classification rule has error at most twice the Bayes error rate.

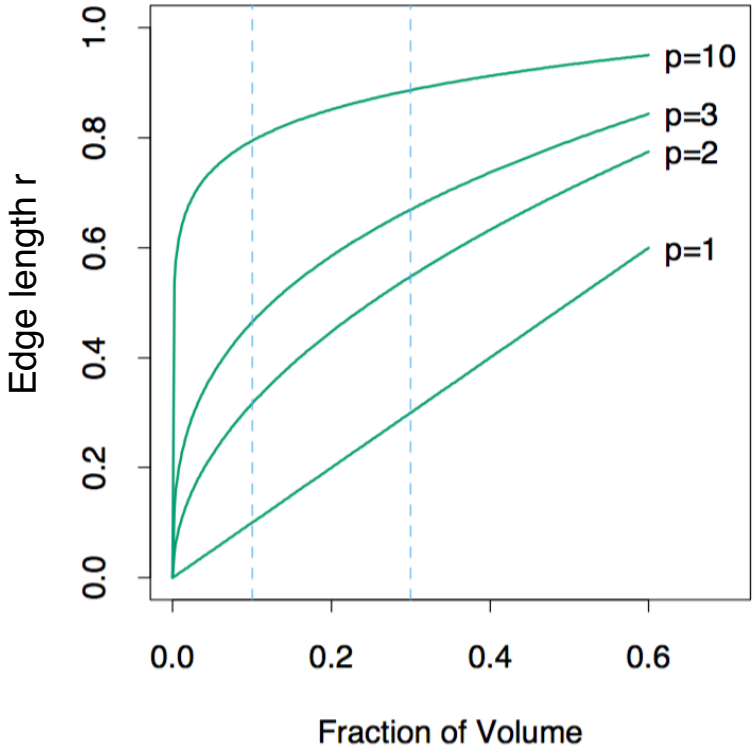
Can fit arbitrarily complicated decision boundaries



Curse of dimensionality, example 1



side length r



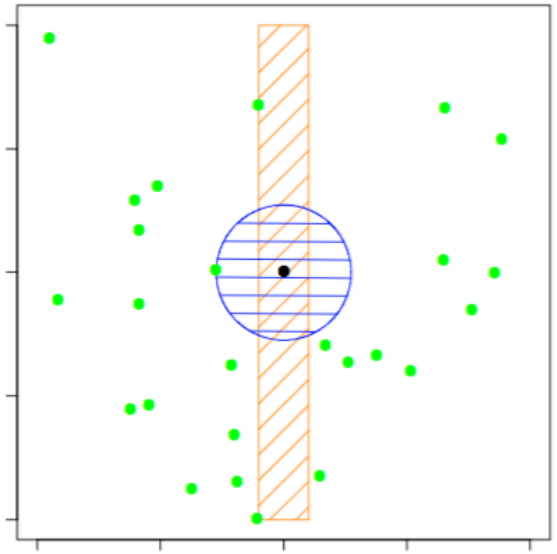
X is uniformly distributed over $[0,1]^p$. What is $\mathbb{P}(X \in [0,r]^p)$?

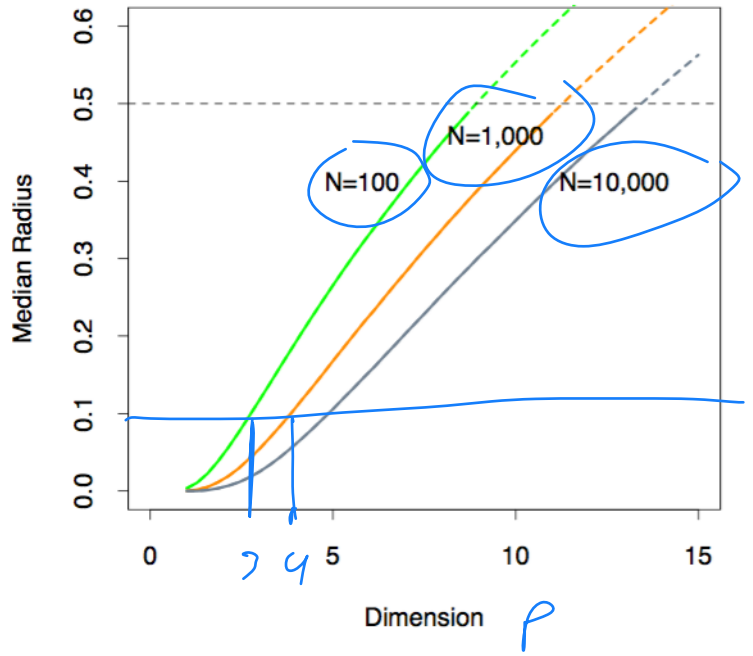
How many samples do we need so that a nearest neighbor is within a cube of side length r?

Curse of dimensionality, example 2

high d space is hard to cover

$\{X_i\}_{i=1}^n$ are uniformly distributed over $[-.5, .5]^p$.

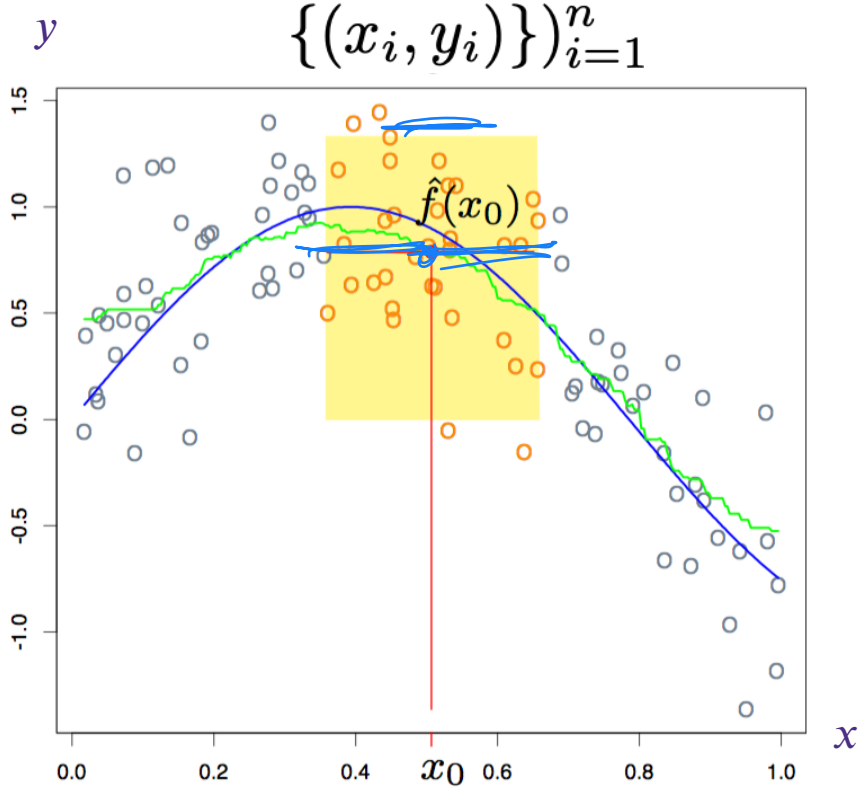




What is the median distance from a point at origin to its 1NN?

How many samples do we need so that a median Euclidean distance is within r?

Nearest neighbor regression



kNN regressor

$$\hat{f}(x) = \frac{1}{k} \sum_{i \in kNN(x)} y^{(i)}$$

$$= \sum_{i=1}^{n} y^{(i)} 1\{x^{(i)} \in kNN(x)\}$$

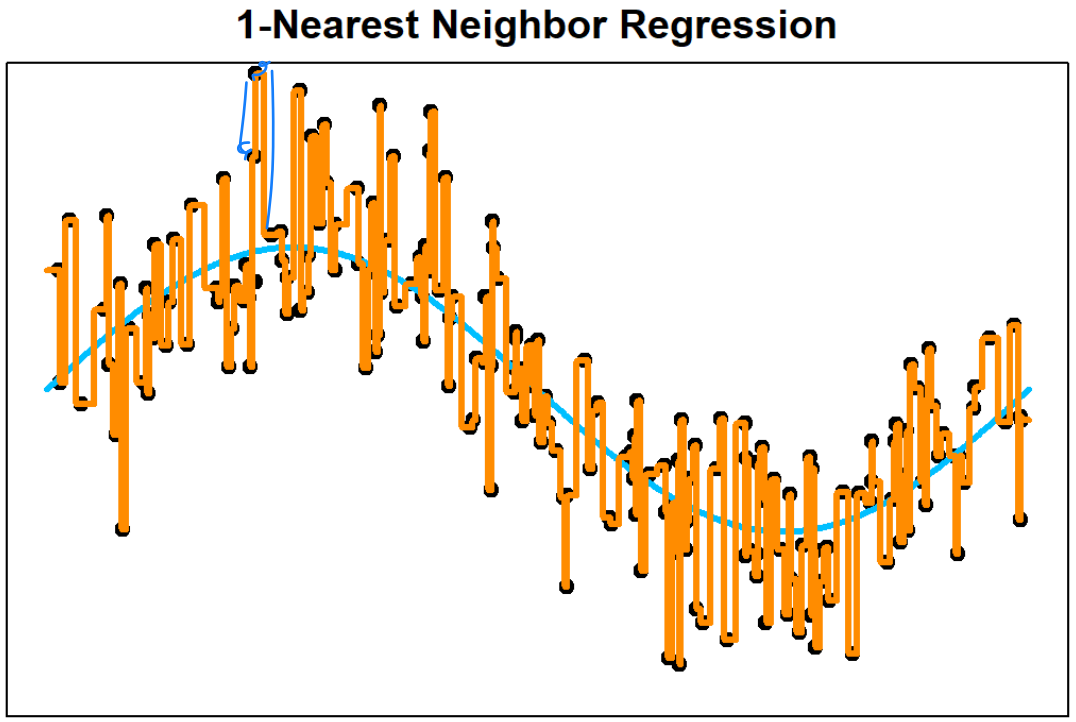
recall

$$f^*(x) = \arg\min_f \mathbb{E}[(f(x) - y)^2]$$

$$= \mathbb{E}[y \mid x]$$

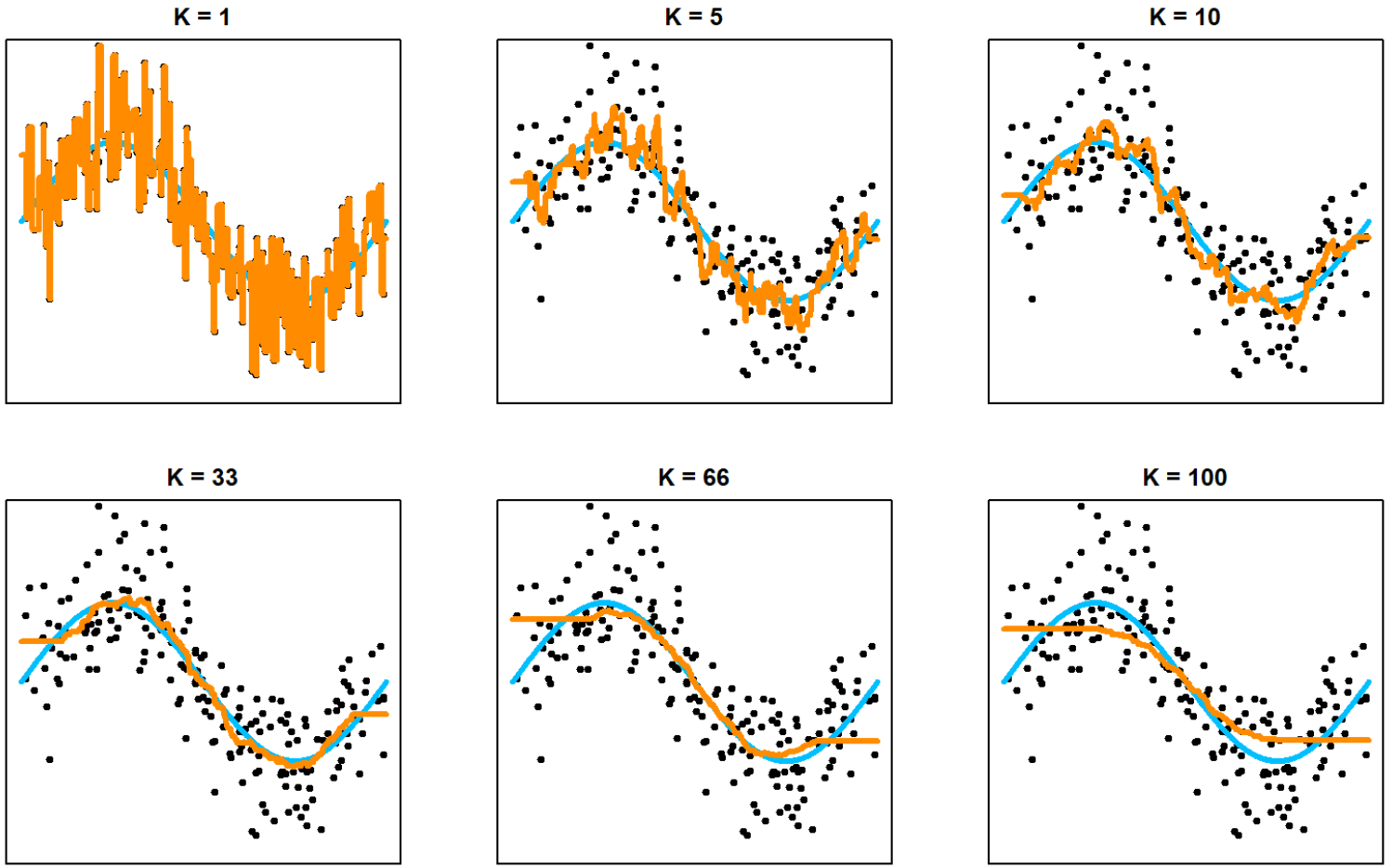
// best in theory

Overfitting



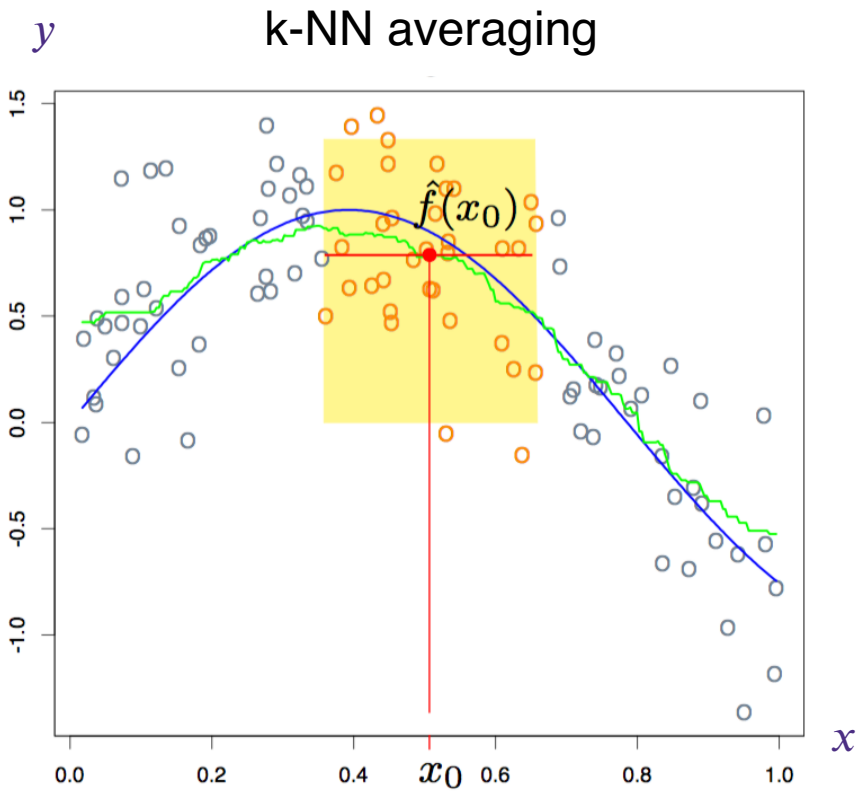
Figures from Ruoqing Zhu, Stat 542

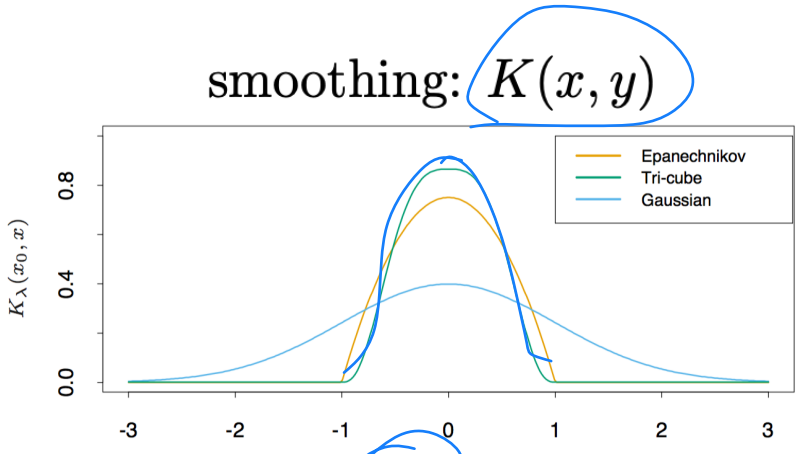
Bias vs variance



Figures from Ruoqing Zhu, Stat 542

Smoothed nearest neighbor regression



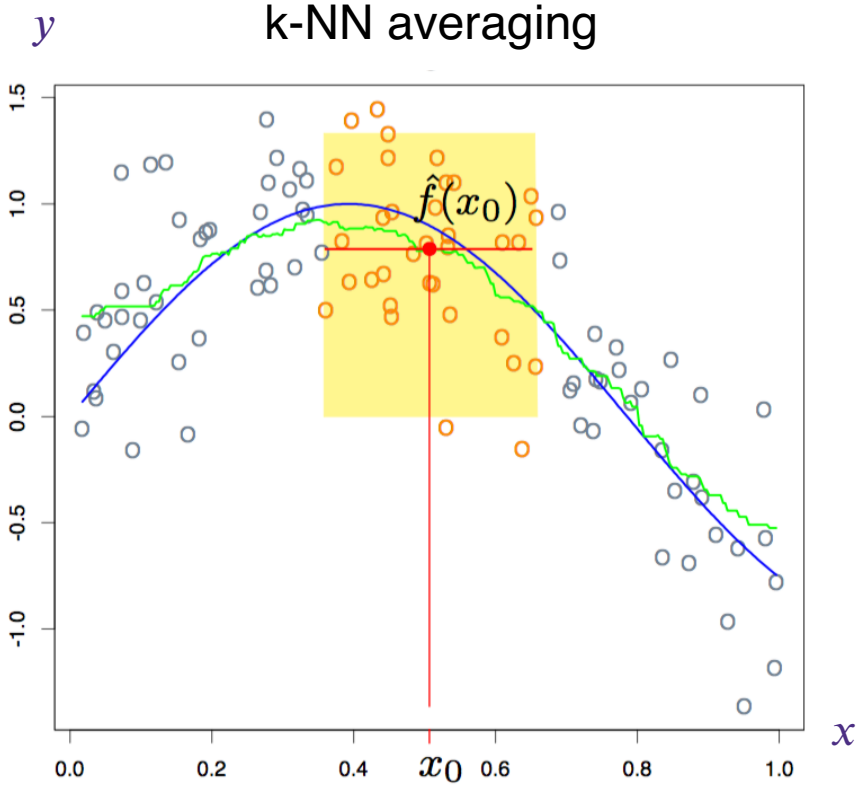


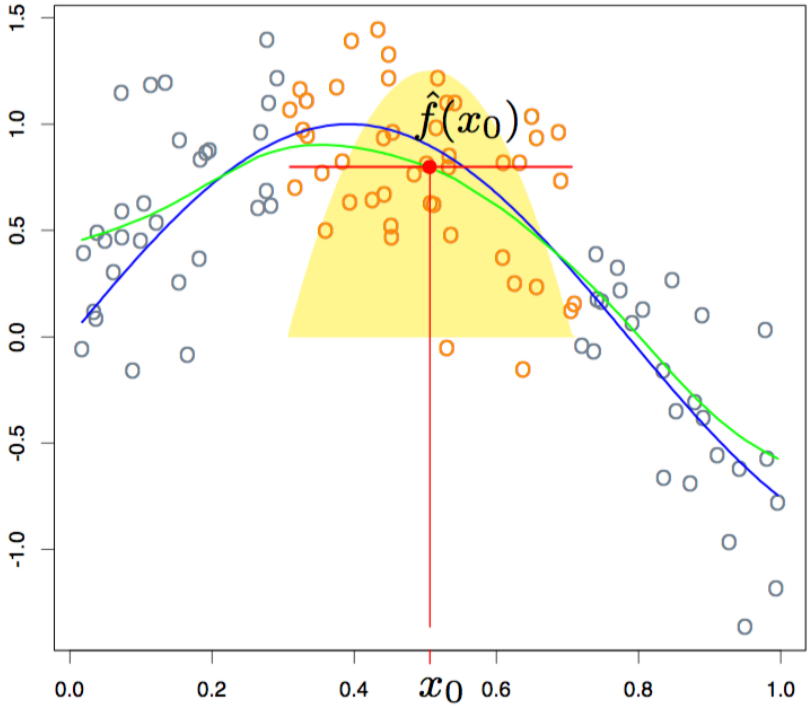
$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(x, x^{(i)}\right) y^{(i)}}{\sum_{i=1}^{n} K\left(x, x^{(i)}\right)}$$

Smoothed nearest neighbor regression

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(x, x^{(i)}\right) y^{(i)}}{\sum_{i=1}^{n} K\left(x, x^{(i)}\right)}$$

k-NN averaging

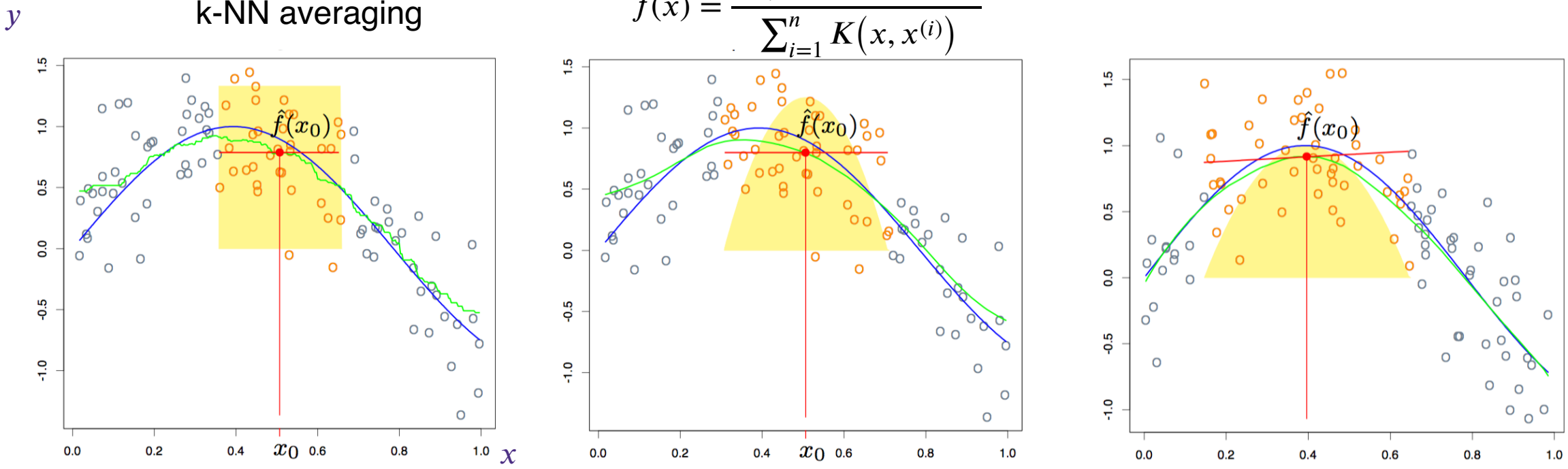




Locally linear regression

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(x, x^{(i)}\right) y^{(i)}}{\sum_{i=1}^{n} K\left(x, x^{(i)}\right)}$$

k-NN averaging



Have we seen non-parametric methods before?

- Kernel methods are non-parametric:

$$f(x) = \sum_{i=1}^{n} \alpha_i K\left(x^{(i)}, x\right)$$

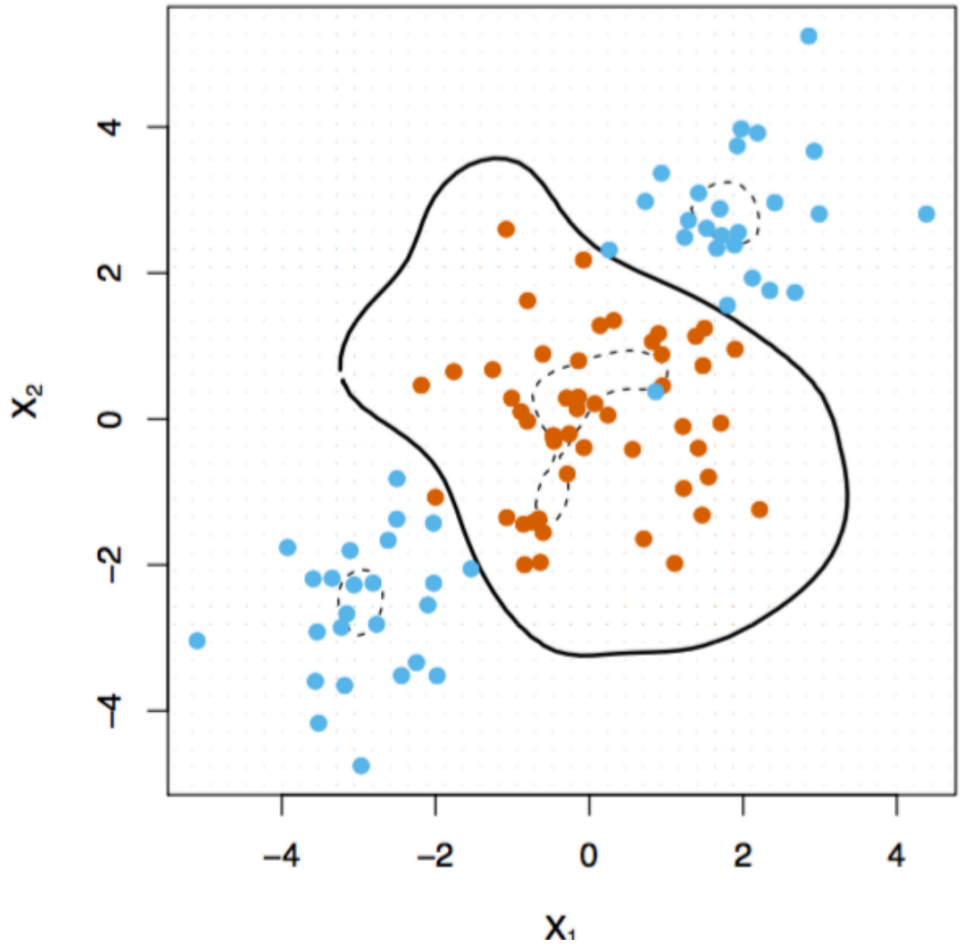
parameters goes up with # data

- Compare with (smoothed) nearest neighbors:

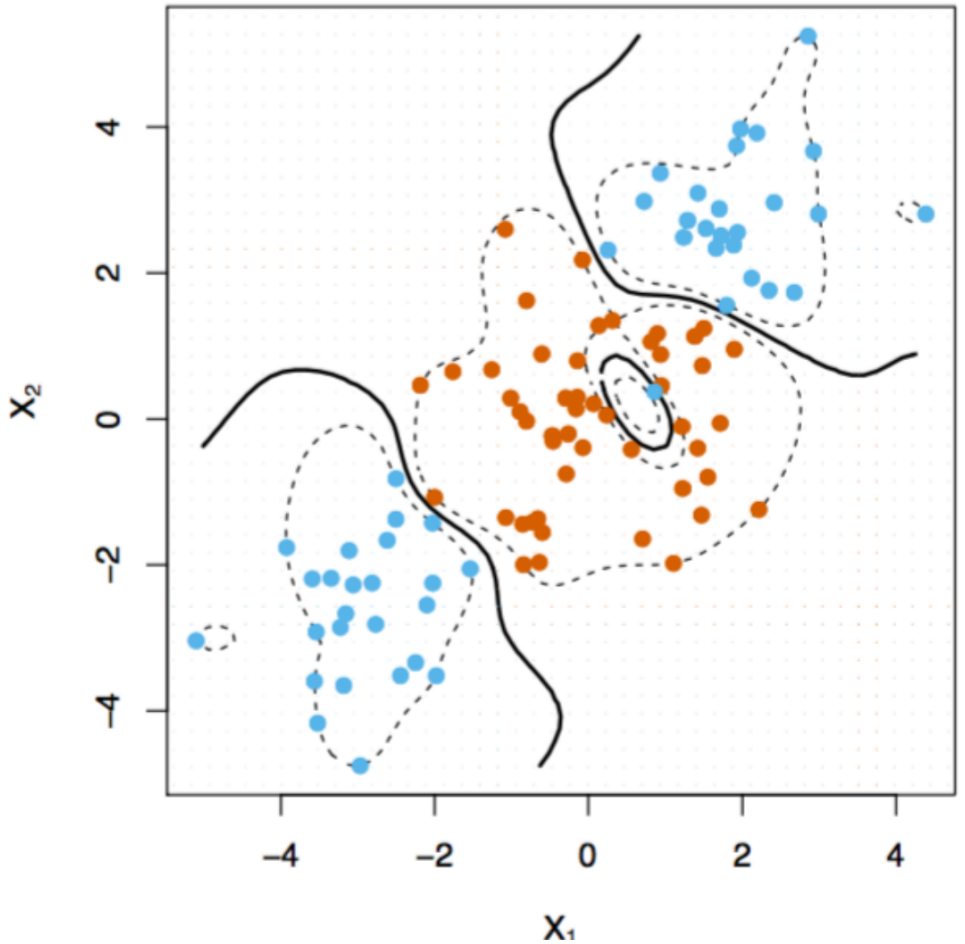
$$\hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(x, x^{(i)}\right) y^{(i)}}{\sum_{i=1}^{n} K\left(x, x^{(i)}\right)}$$

The Radial Basis Function (RBF) kernel $\exp\left(-\frac{\|x^{(i)} - x\|_2^2}{2\sigma^2}\right)$

Bandwidth σ is large enough

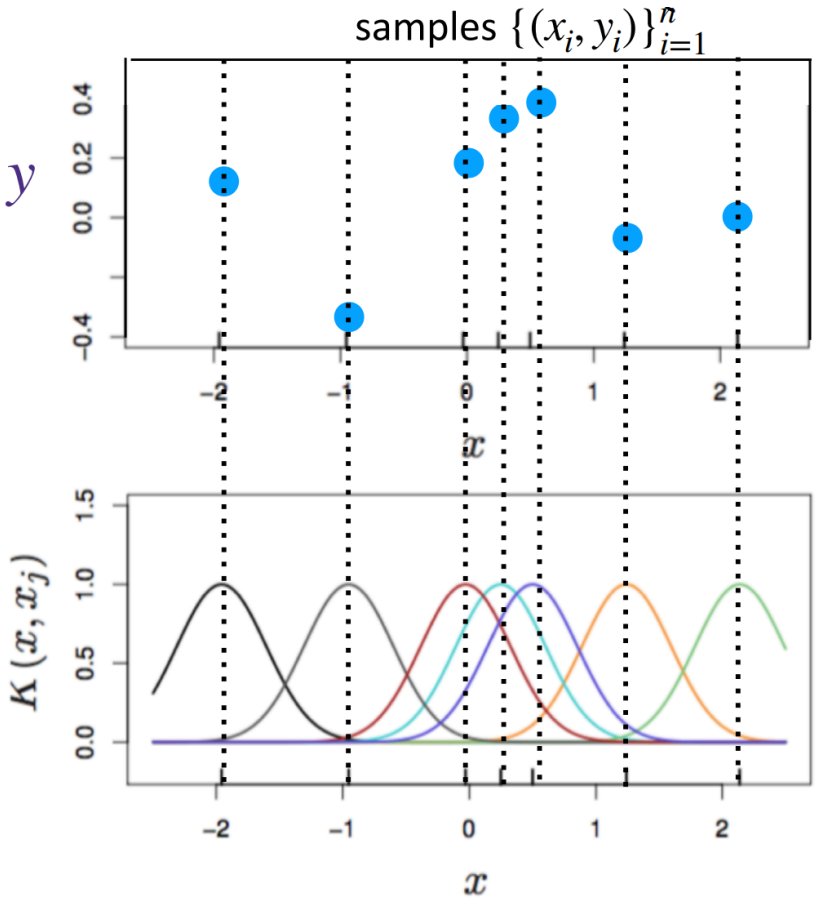


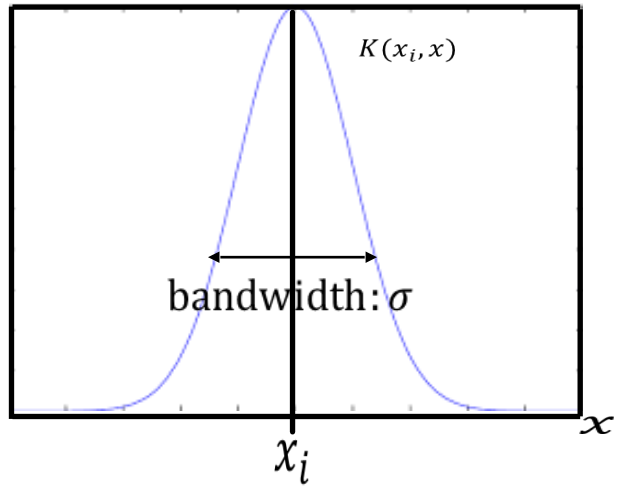
Bandwidth σ is small



The Radial Basis Function (RBF) kernel $\exp\left(-\frac{\left\|x^{(i)}-x\right\|_2^2}{2\sigma^2}\right)$

$$f(x) = \sum_{i=1}^{n} \alpha_i K\left(x^{(i)}, x\right)$$





Kernel methods

$$f(x) = \sum_{i=1}^{n} \alpha_i K\left(x^{(i)}, x\right)$$

- Can be seen as a soft, learned version of "nearest" neighbors
- $K\left(x^{(i)}, x\right) = \phi\left(x^{(i)}\right)^\top \phi(x)$ defines "similarity" between $x^{(i)}$ and x
- How many parameters?

Takeaways

- k-NN is very simple to explain and implement
- No training! But inference can still be computationally demanding.
- You can use other forms of distance (not just Euclidean)
- Smoothing and local linear regression can improve performance (at the cost of higher variance)
- With a lot of data, “local methods” have strong, simple theoretical guarantees
- Without a lot of data, neighborhoods aren’t “local” and methods suffer (curse of dimensionality)