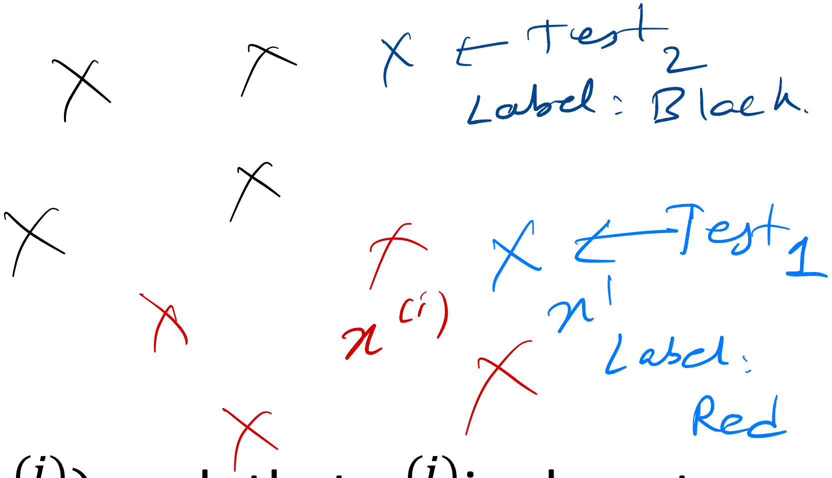
CS 4375
Nearest Neighbor Methods

Rishabh Iyer
University of Texas at Dallas

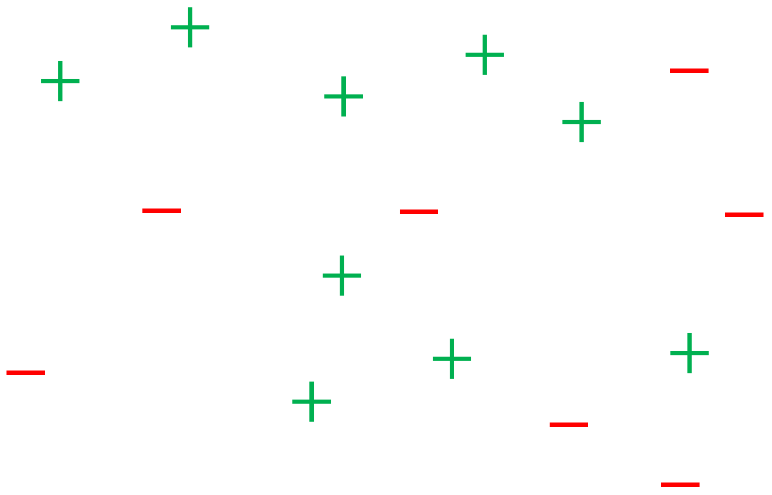
Based on the slides of Vibhav Gogate, Nick Rouzzi, David Sontag and few other sources

Nearest Neighbor Methods

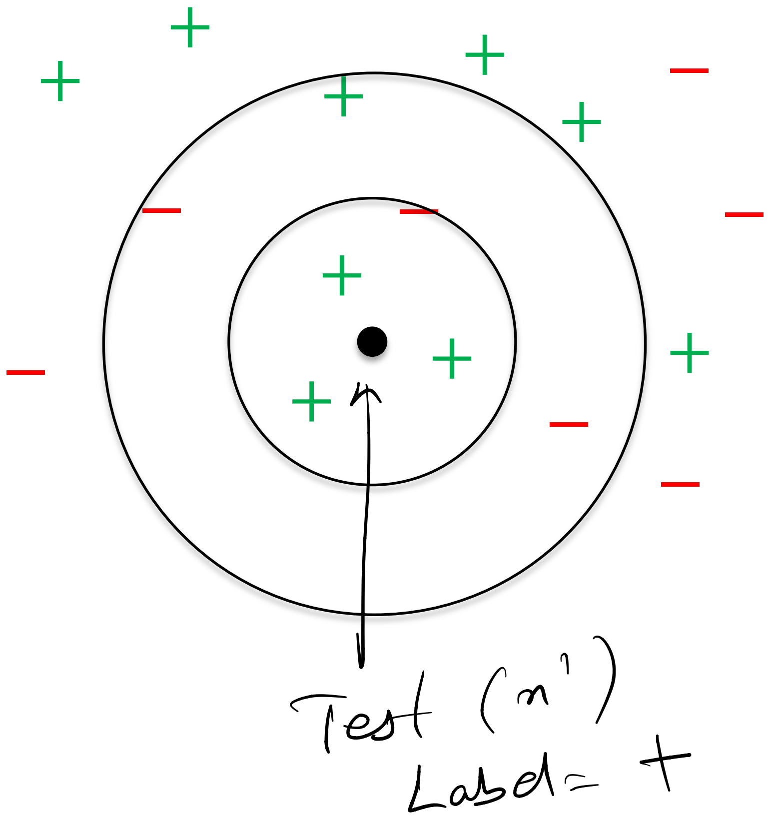


- Learning
  - Store all training examples
- Classifying a new point x'
  - Find the training example $(x^{(i)}, y^{(i)})$ such that $x^{(i)}$is closest (for some notion of close) to x'
  - Classify x' with the label $y^{(i)}$

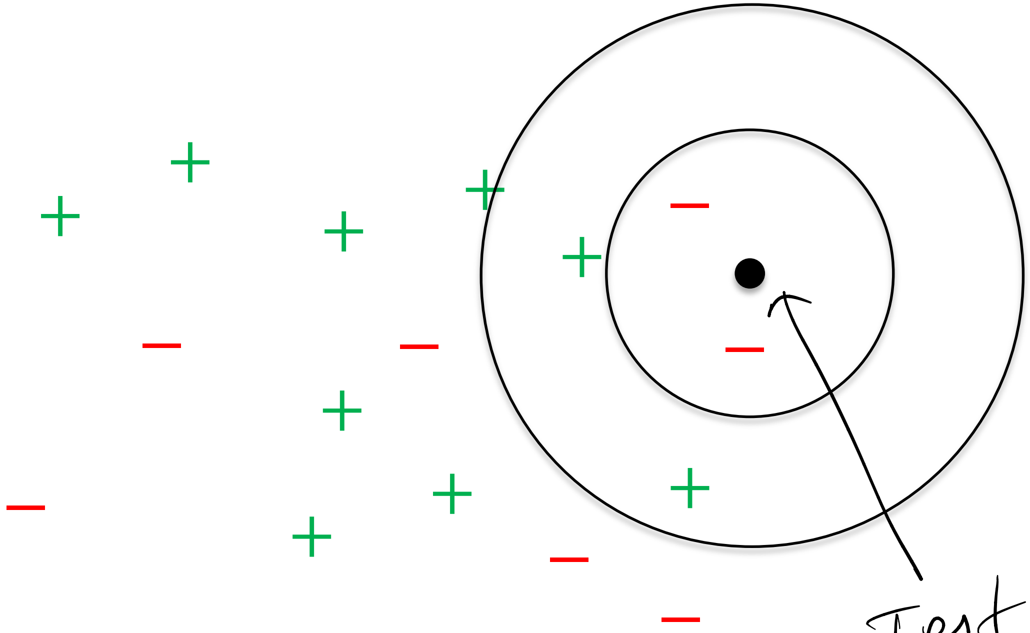
Nearest Neighbor Methods



Nearest Neighbor Methods



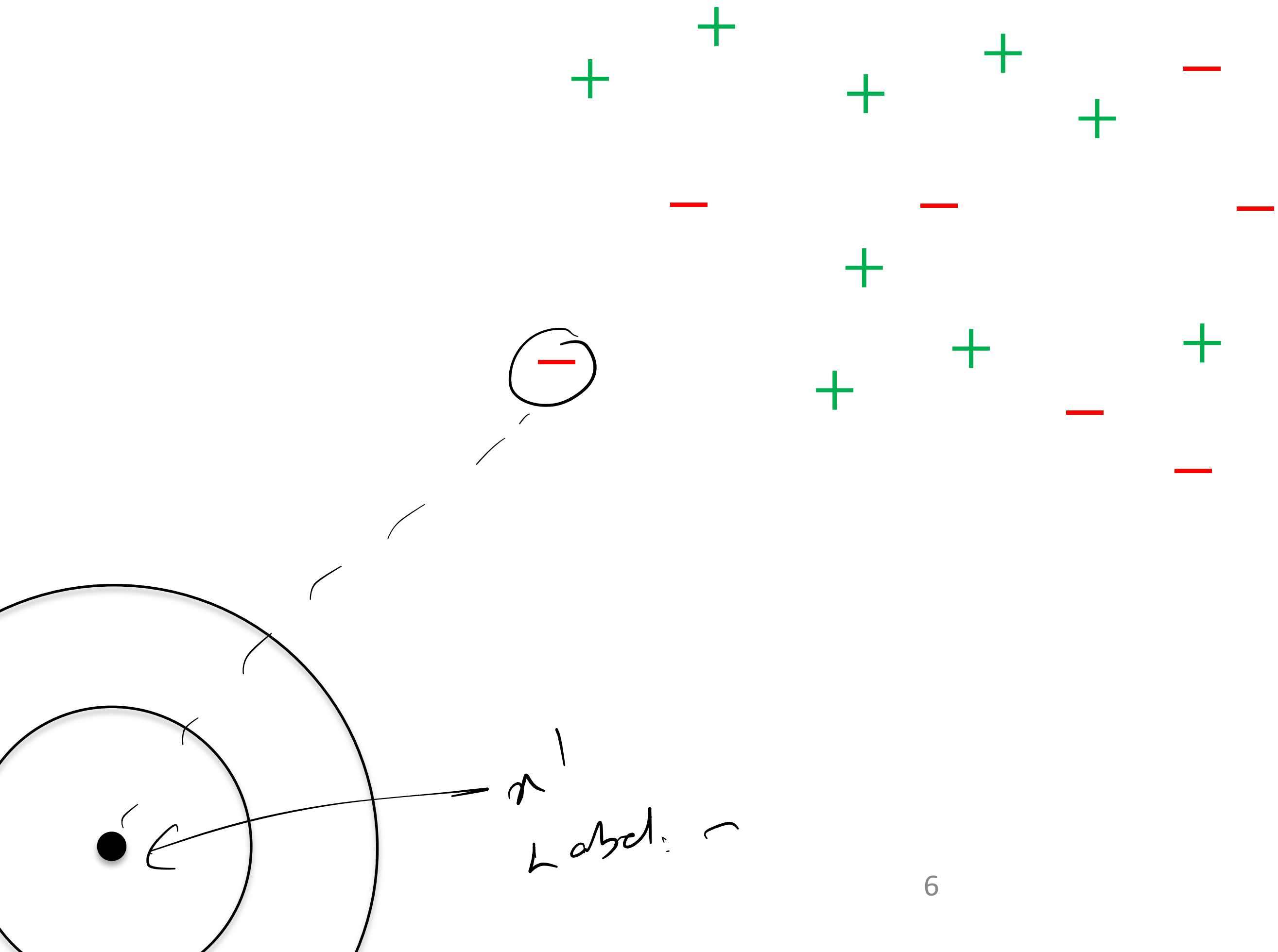
Nearest Neighbor Methods



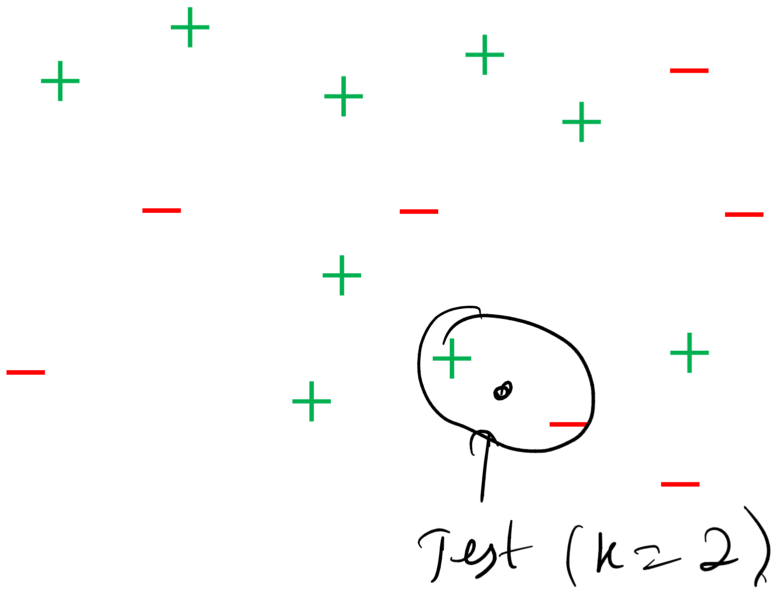
Test (x')

Label: −

Nearest Neighbor Methods

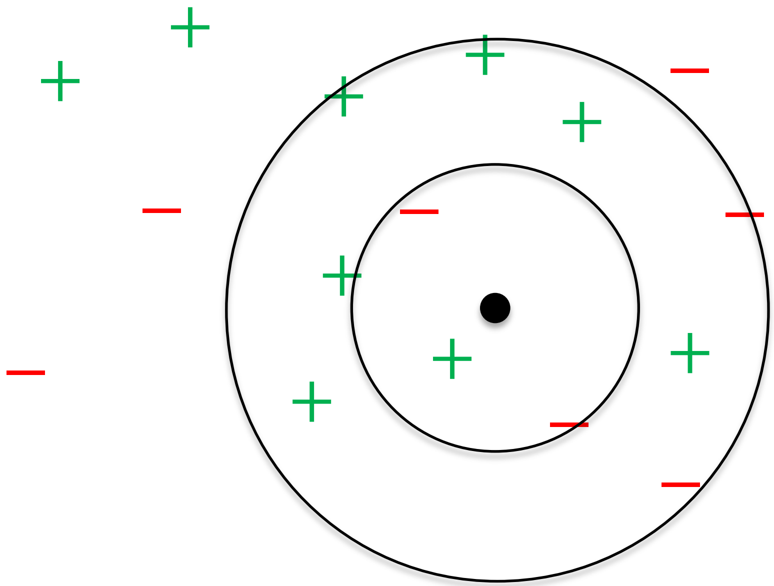


Nearest Neighbor Methods



k-nearest neighbor methods look at the k closest points in the training set and take a majority vote (should choose k to be odd)

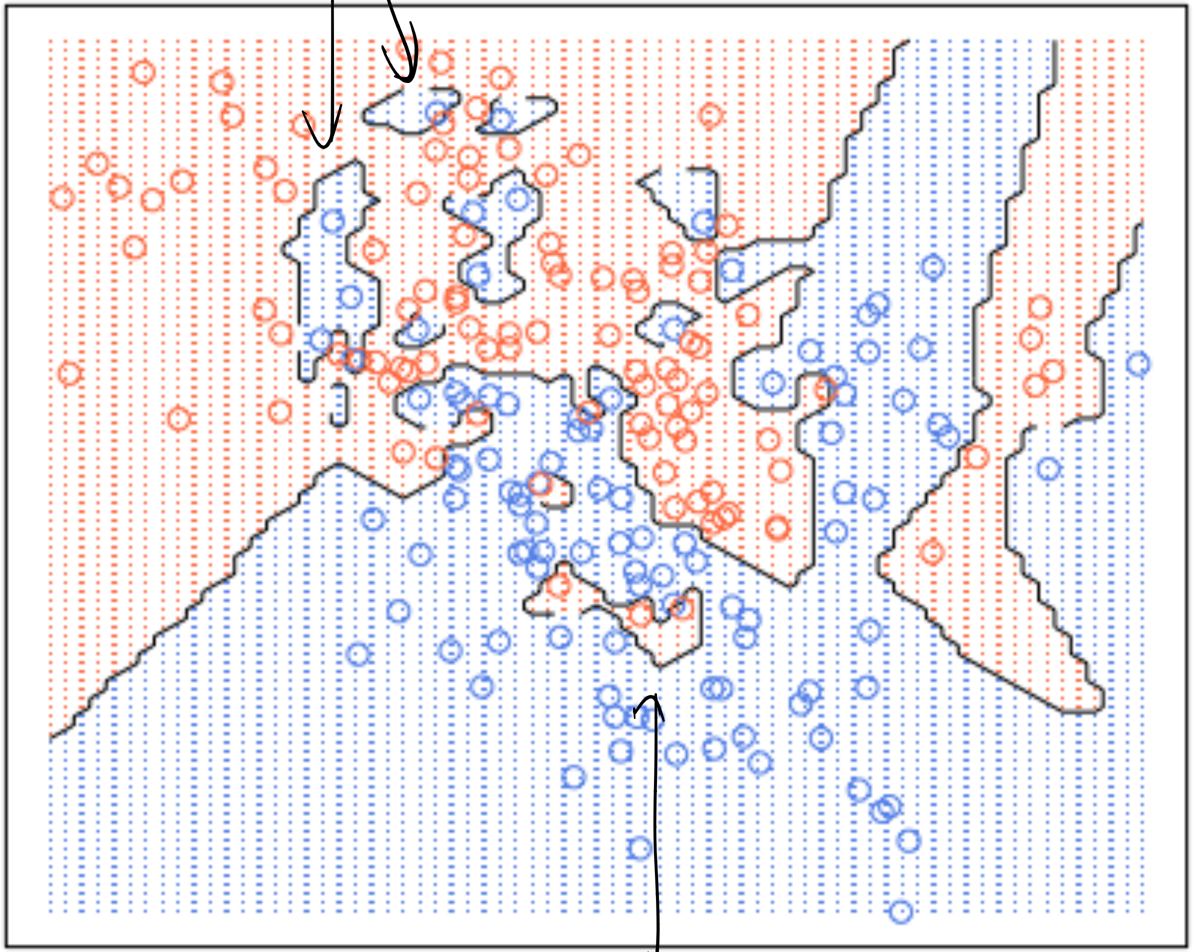
Nearest Neighbor Methods



k-nearest neighbor methods look at the k closest points in the training set and take a majority vote (should choose k to be odd)

1-NN Example

Islands

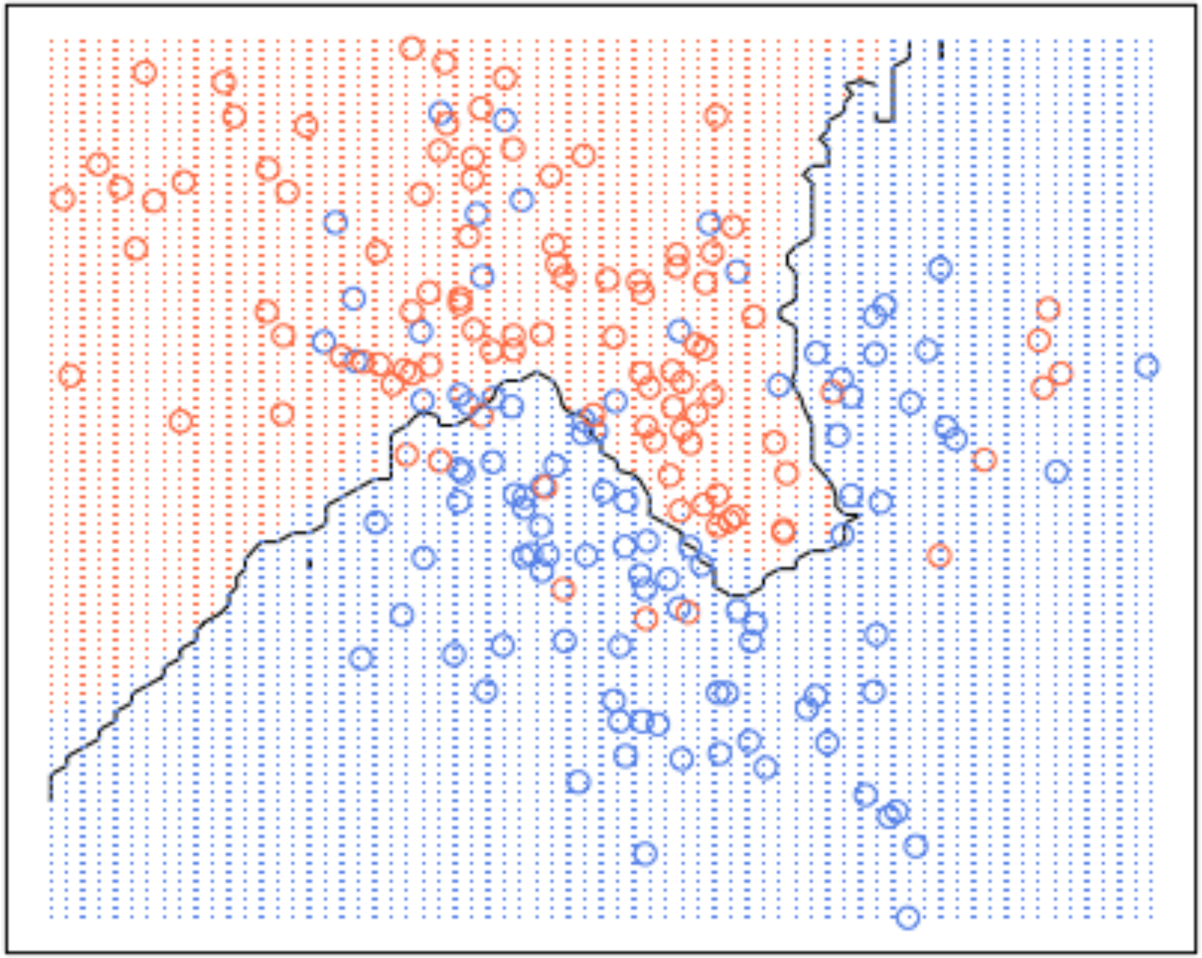


Islands

Looks like overfitting.

[Kevin Zakka]

20-NN Example



Better Fit

Maybe a little underfit.

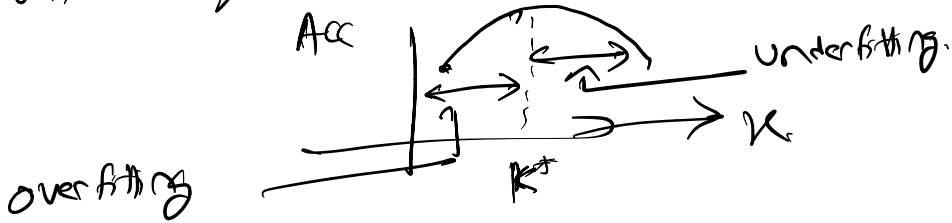
[Kevin Zakka]

K = # Nearest Neighbors

<u>Hyperparameter</u>

How to find optimal K?

1. Divide Dataset into (Train, Val, Test)
2. Find a range of "K" (for e.g 1:2:100)
3. Train KNN model for each K on Train & eval on Val set
4. Find K with highest Val Acc



Nearest Neighbor Methods

- Applies to data sets with points in $\mathbb{R}^d$
  - Best for large data sets with only a few (< 20) attributes

(M = Large) (features)

- Advantages
  - Learning is easy
  - Can learn complicated decision boundaries
- Disadvantages
  - Classification is slow (need to keep the entire training set around) ← High memory requirement.
  - Easily fooled by irrelevant attributes

Practical Challenges

- How to choose the right measure of closeness?
  - Euclidean distance is popular, but many other possibilities
- How to pick k?
  - Too small and the estimates are noisy, too large and the accuracy suffers
- What if the nearest neighbor is really far away?

$$D_2(x,y) = \|x-y\|_2 = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2+ \ldots (x_d-y_d)^2}$$

$$D_1(x,y) = \|x-y\|_1 = |x_1-y_1| + |x_2-y_2| + \ldots + |x_d-y_d|$$

Choosing the Distance

- Euclidean distance makes sense when each of the features is roughly on the same scale

  feet 1-100

  - If the features are very different (e.g., height and age), then Euclidean distance makes less sense as height would be less significant than age simply because age has a larger range of possible values
  - To correct for this, feature vectors are often recentered around their means and scaled by the standard deviation over the training set

↑

Normalization.

Normalization

$$\hat{x}_i = \frac{x_i - \overline{x_i}}{\sigma_i}$$

- Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x^{(i)}$$

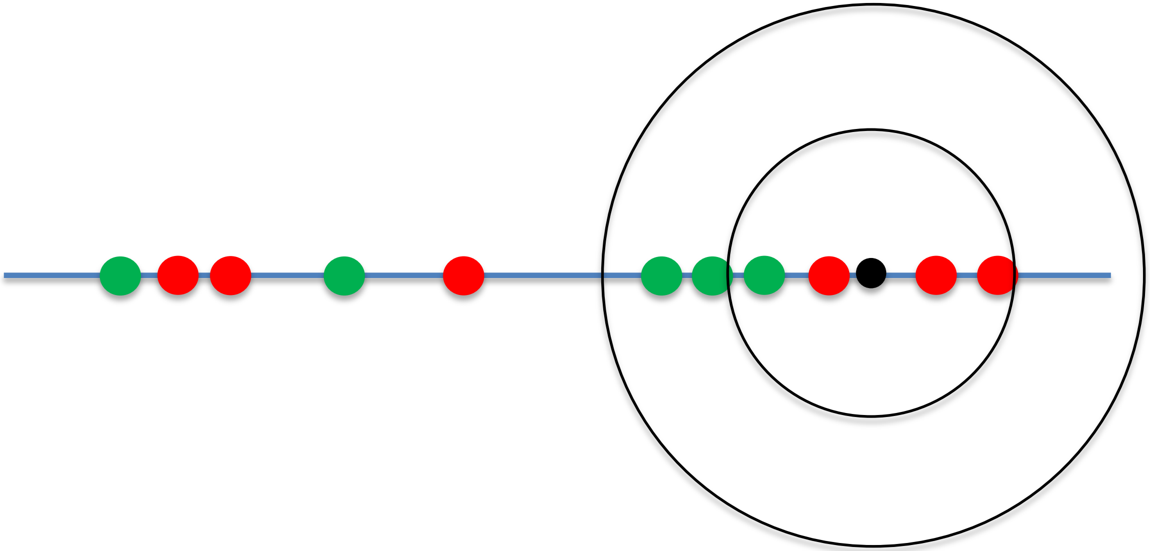
- Sample variance (biased)

$$\hat{\sigma}_k^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_k^{(i)} - \bar{x}_k\right)^2$$

$$\hat{x}_i = \frac{x_i - x_i^{min}}{x_i^{max} - x_i^{min}} \in [0, 1]$$

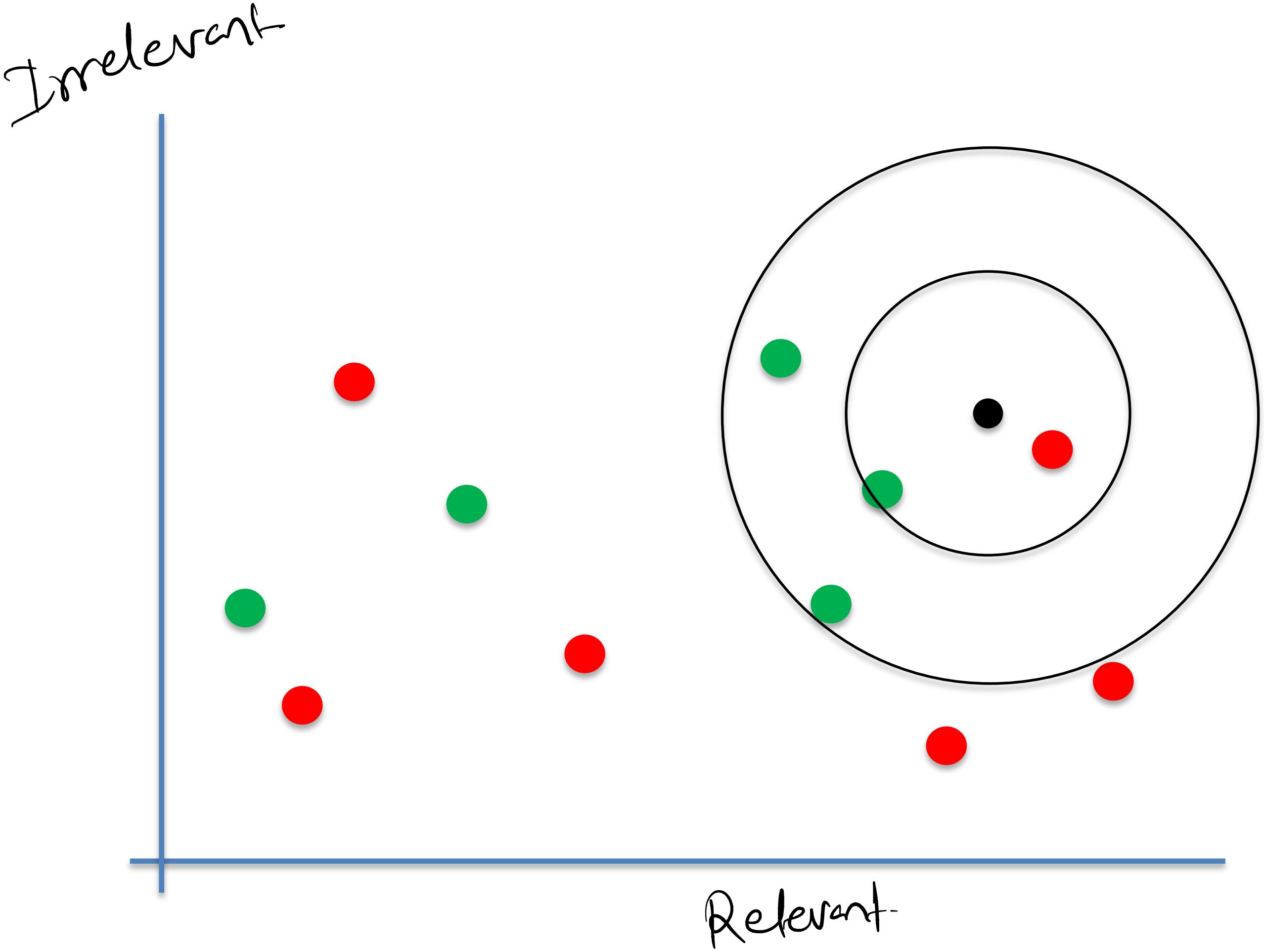
Irrelevant Attributes

Consider the nearest neighbor problem in one dimension



Irrelevant Attributes

Now, add a new attribute that is just random noise...



K-Dimensional Trees

- In order to do classification, we can compute the distances between all points in the training set and the point we are trying to classify

100k 20

  - With m data points in n-dimensional space, this takes $O(mn)$ time for Euclidean distance
  - It is possible to do better if we do some preprocessing on the training data

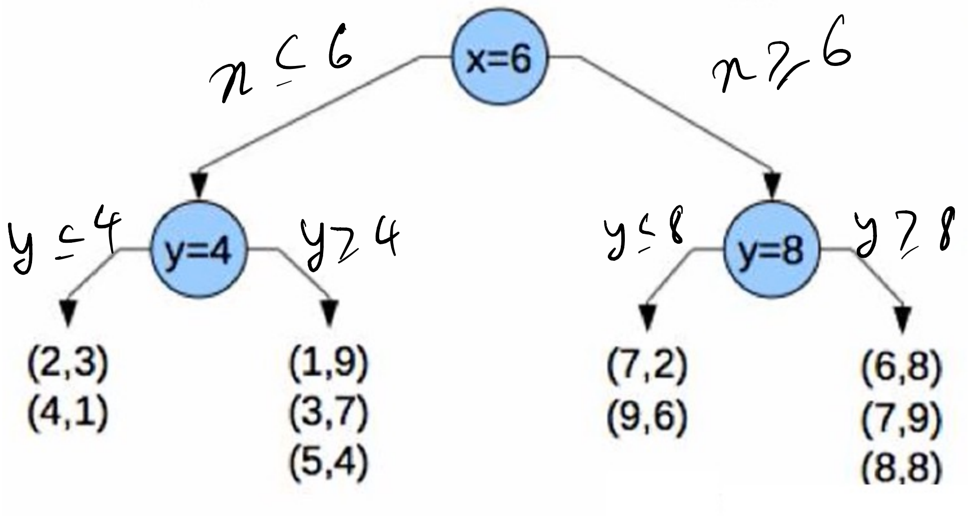
$O(\underline{mn} + m\log m)$

K-Dimensional Trees

- k-d trees provide a data structure that can help simplify the classification task by constructing a tree that partitions the search space
  - Starting with the entire training set, choose some dimension, i
  - Select an element of the training data whose i^{th} dimension has the median value among all elements of the training set
  - Divide the training set into two pieces: depending on whether their i^{th} attribute is smaller or larger than the median
  - Repeat this partitioning process on each of the two new pieces separately

K-Dimensional Trees

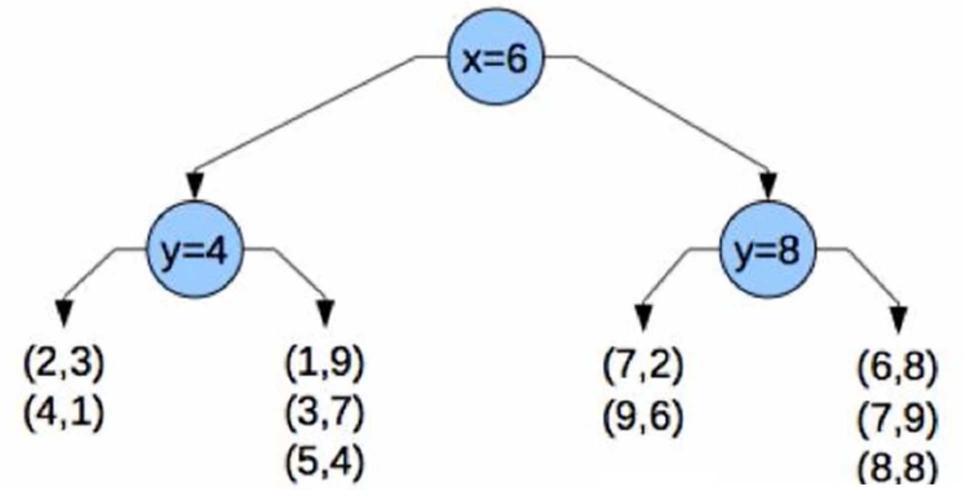
- Building a K-D tree from training data:
  - {(1,9), (2,3), (4,1), (3,7), (5,4), (6,8), (7,2), (8,8), (7,9), (9,6)}
  - pick random dimension, find median, split data, repeat

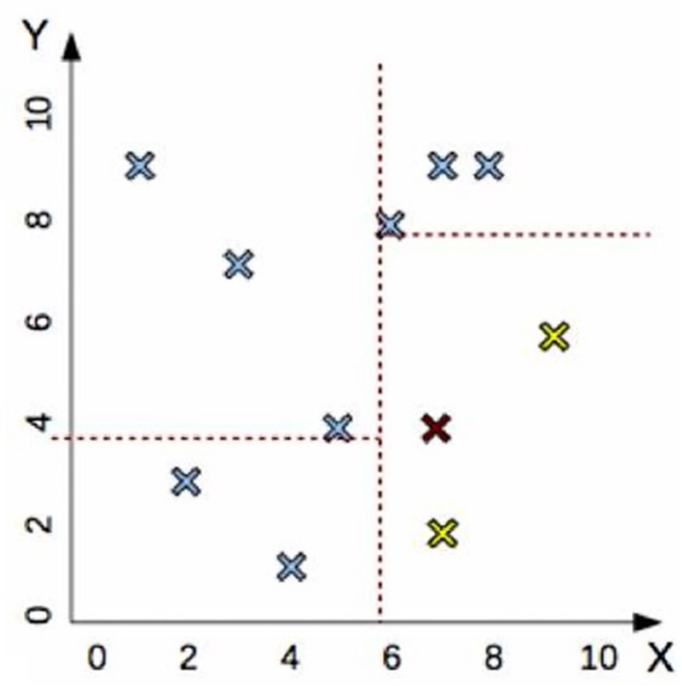


Adapted from Victor Lavrenko

K-Dimensional Trees

- Building a K-D tree from training data:
  - {(1,9), (2,3), (4,1), (3,7), (5,4), (6,8), (7,2), (8,8), (7,9), (9,6)}
  - pick random dimension, find median, split data, repeat

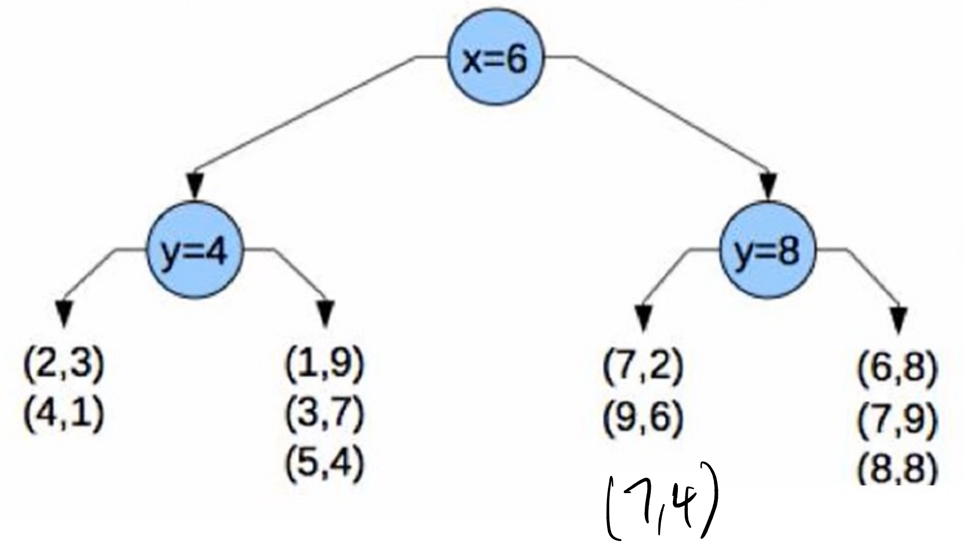


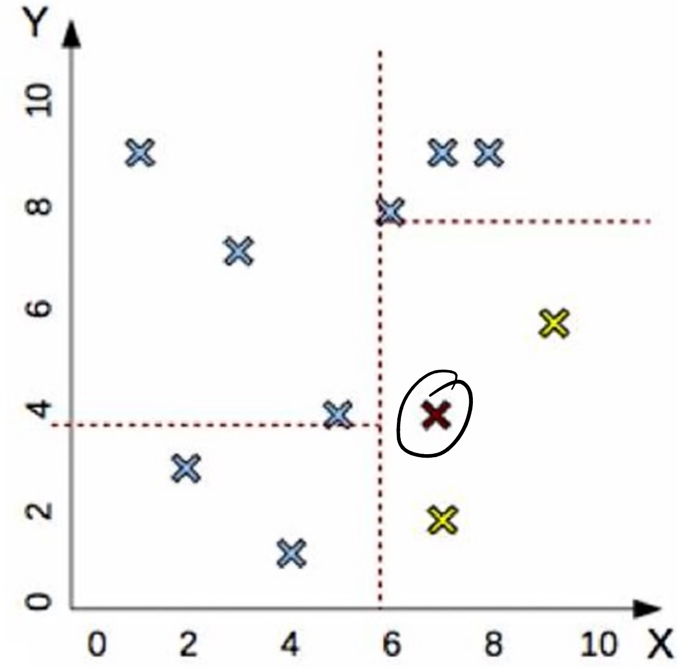


Adapted from Victor Lavrenko

K-Dimensional Trees: Inference

- Find NNs for new point (7,4)
  - find region containing (7,4)
  - compare to all points in region





Adapted from Victor Lavrenko

K-Dimensional Trees

- By design, the constructed k-d tree is “bushy”
  - The idea is that if new points to classify are evenly distributed throughout the space, then the expected (amortized) cost of classification is approximately $O(d \log n)$ operations
- Summary
  - k-NN is fast and easy to implement
  - No training required
  - Can be good in practice (where applicable)