



10-301/10-601 Introduction to Machine Learning

Machine Learning Department
School of Computer Science
Carnegie Mellon University

Overfitting + k-Nearest Neighbors

Matt Gormley
Lecture 4
Sep. 11, 2023



Course Staff

Team A (HW2, HW6)



Alisa Qiu

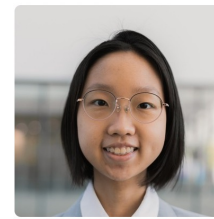


Annie Wu



Bhargav Hadya

Team B (HW3, HW7)



Haohui Liu



Monica Geng



Sahithya Senthilkumar

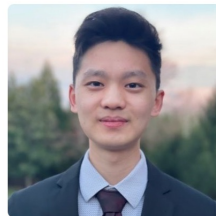
Education Associate



Brynn Edmunds



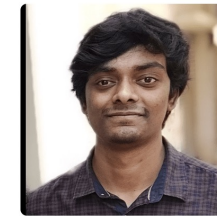
Erin Gao



Sebastian Lu

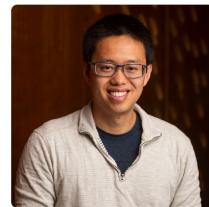


Sivaramakrishnan
Subramanian



Rakshith Srinivasa Murthy

Instructors

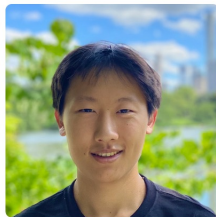


Henry Chai



Matt Gormley

Team C (HW4, HW8)



Kevin Ren

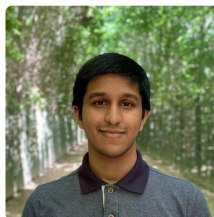


Meher Mankikar



Pranit Chawla

Team D (HW5, HW9)



Abhishek Vijayakumar



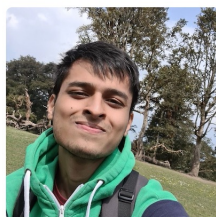
Ally Du



Andrew Wang



Tanvi Karandikar



Yash Gupta



Emily Xie



Neelansh Kaabra

Q&A

Q: Why don't my entropy calculations match those on the slides?

A: Remember that $H(Y)$ is conventionally reported in “bits” and computed using log base 2.
e.g., $H(Y) = -P(Y=0) \log_2 P(Y=0) - P(Y=1) \log_2 P(Y=1)$

Q: When and how do we decide to stop growing trees? What if the set of values an attribute could take was really large or even infinite?

A: We'll address this question for discrete attributes today. If an attribute is real-valued, there's a clever trick that only considers $O(L)$ splits where $L = \#$ of values the attribute takes in the training set. Can you guess what it does?

Q&A

Q: What does decision tree training do if a branch receives no data?

A: Then we hit the base case and create a leaf node. So the real question is what does majority vote do when there is no data? Of course, there is no majority label, so (if forced to) we could just return one randomly.

Q: What do we do at test time when we observe a value for a feature that we didn't see at training time.

A: This really just a variant of the first question. That said, a real DT implementation needs to elegantly handle this case. We could do so by either (a) assuming that all possible values will be seen at train time, so there should be a branch for all attributes even if the partition of the dataset doesn't include them all or (b) recognize the unseen value at test time and return some appropriate label in that case.

Reminders

- **Homework 2: Decision Trees**
 - **Out: Wed, Sep. 6**
 - **Due: Fri, Sep. 15 at 11:59pm**

EMPIRICAL COMPARISON OF SPLITTING CRITERIA

Experiments: Splitting Criteria

Bluntine & Niblett (1992) compared 4 criteria (random, Gini, mutual information, Marshall) on 12 datasets

Medical Diagnosis Datasets: (4 of 12)

- **hypo:** data set of 3772 examples records expert opinion on possible hypo- thyroid conditions from 29 real and discrete attributes of the patient such as sex, age, taking of relevant drugs, and hormone readings taken from drug samples.
- **breast:** The classes are reoccurrence or non-reoccurrence of breast cancer sometime after an operation. There are nine attributes giving details about the original cancer nodes, position on the breast, and age, with multi-valued discrete and real values.
- **tumor:** examples of the location of a primary tumor
- **lymph:** from the lymphography domain in oncology. The classes are normal, metastases, malignant, and fibrosis, and there are nineteen attributes giving details about the lymphatics and lymph nodes

Table 1. Properties of the data sets

Data Set	Classes	Attr.s	Training Set	Test Set
hypo	4	29	1000	2772
breast	2	9	200	86
tumor	22	18	237	102
lymph	4	18	103	45
LED	10	7	200	1800
mush	2	22	200	7924
votes	2	17	200	235
votes1	2	16	200	235
iris	3	4	100	50
glass	7	9	100	114
xd6	2	10	200	400
pole	2	4	200	1647

Experiments: Splitting Criteria

Table 3. Error for different splitting rules (pruned trees).

Data Set	Splitting Rule			
	GINI	Info. Gain	Marsh.	Random
hypo	1.01 ± 0.29	0.95 ± 0.22	1.27 ± 0.47	7.44 ± 0.53
breast	28.66 ± 3.87	28.49 ± 4.28	27.15 ± 4.22	29.65 ± 4.97
tumor	60.88 ± 5.44	62.70 ± 3.89	61.62 ± 3.98	67.94 ± 5.68
lymph	24.44 ± 6.92	24.00 ± 6.87	24.33 ± 5.51	32.33 ± 11.25
LED	33.77 ± 3.06	32.89 ± 2.59	33.15 ± 4.02	38.18 ± 4.57
mush	1.44 ± 0.47	1.44 ± 0.47	7.31 ± 2.25	8.77 ± 4.65
votes	4.47 ± 0.95	4.57 ± 0.87	11.77 ± 3.95	12.40 ± 4.56
votes1	12.79 ± 1.48	13.04 ± 1.65	15.13 ± 2.89	15.62 ± 2.73
iris	5.00 ± 3.08	4.90 ± 3.08	5.50 ± 2.59	14.20 ± 6.77
glass	39.56 ± 6.20	50.57 ± 6.73	40.53 ± 6.41	53.20 ± 5.01
xd6	22.14 ± 3.23	22.17 ± 3.36	22.06 ± 3.37	31.86 ± 3.62
pole	15.43 ± 1.51	15.47 ± 0.88		

Key Takeaway:
GINI gain and
Mutual
Information are
statistically
indistinguishable!



Info. Gain is another name
for *mutual information*

Experiments: Splitting Criteria

Table 4. Difference and significance of error for GINI splitting rule versus others.

Data Set	Splitting Rule		
	Info. Gain	Marsh.	Random
hypo	-0.06 (0.82)	0.26 (0.99)	6.43 (1.00)
breast	-0.17 (0.23)	-1.51 (0.94)	0.99 (0.72)
tumor	1.81 (0.84)	0.74 (0.39)	7.06 (0.99)
lymph	-0.44 (0.83)	0.11 (0.05)	7.89 (0.99)
LED	0.12 (0.17)	6.58	
mush	0.00 (0.00)	5.86	
votes	0.11 (0.55)	7.30	
votes1	0.26 (0.47)	2.34	
iris	-0.10 (0.67)	0.50	
glass	1.01 (0.50)	0.96	
xd6	0.04 (0.11)	-0.07	
pole	0.03 (0.11)	-0.43	

Key Takeaway:
GINI gain and Mutual Information are statistically indistinguishable!



Results are of the form A.AA (B.BB) where:

1. A.AA is the **average difference in errors** between the two methods
2. B.BB is the **significance** of the difference according to a two-tailed **paired t-test**

INDUCTIVE BIAS (FOR DECISION TREES)

Decision Tree Learning Example

Dataset:

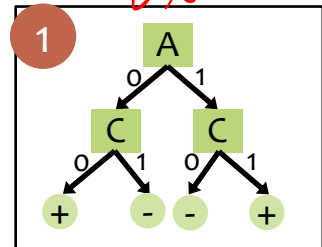
Output Y, Attributes A, B, C

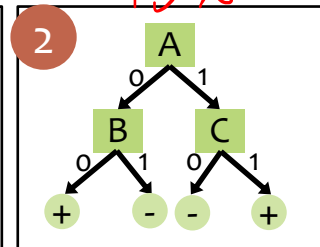
Y	A	B	C
+	0	0	0
+	0	0	1
-	0	1	0
+	0	1	1
-	1	0	0
-	1	0	1
-	1	1	0
+	1	1	1

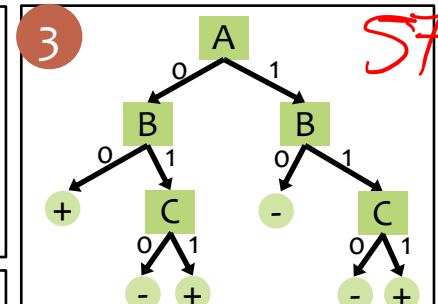
In-Class Exercise

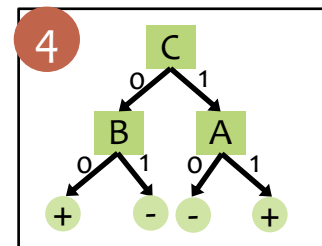
Which of the following trees would be learned by the decision tree learning algorithm using “error rate” as the splitting criterion?

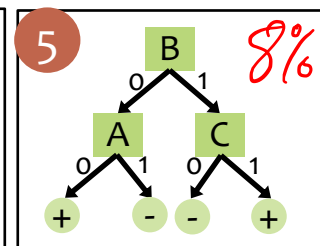
(Assume ties are broken alphabetically.)

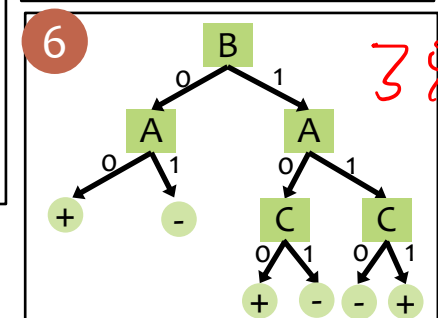
1 15% 

2 15% 

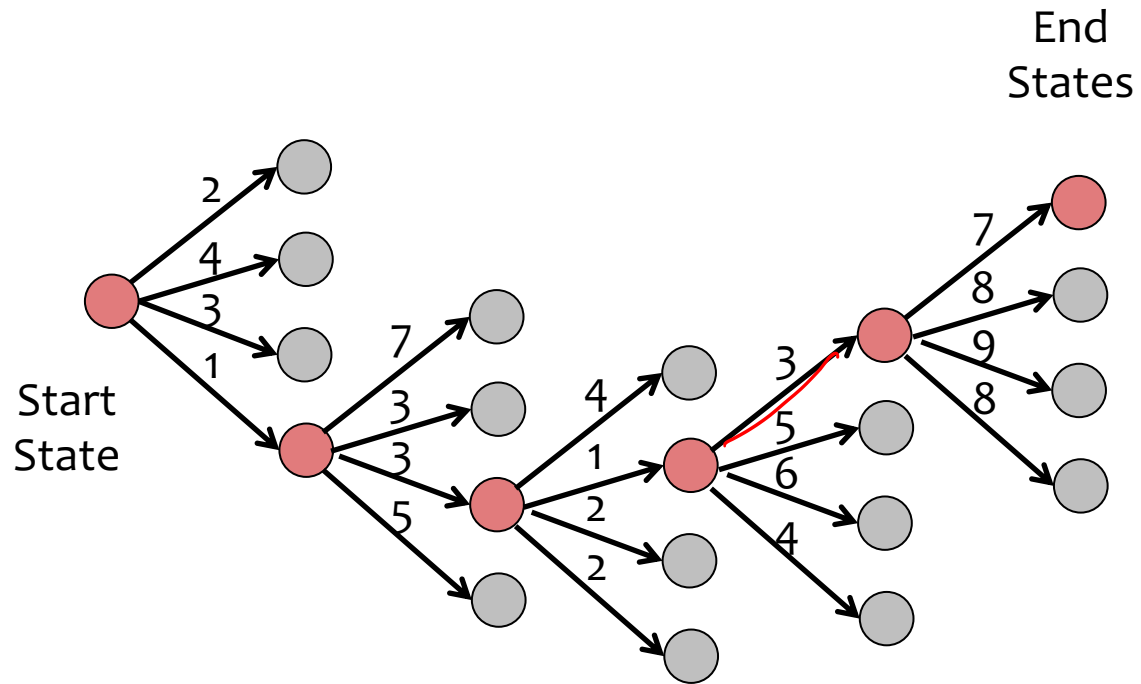
3 57% 

4 

5 8% 

6 3% 

Background: Greedy Search



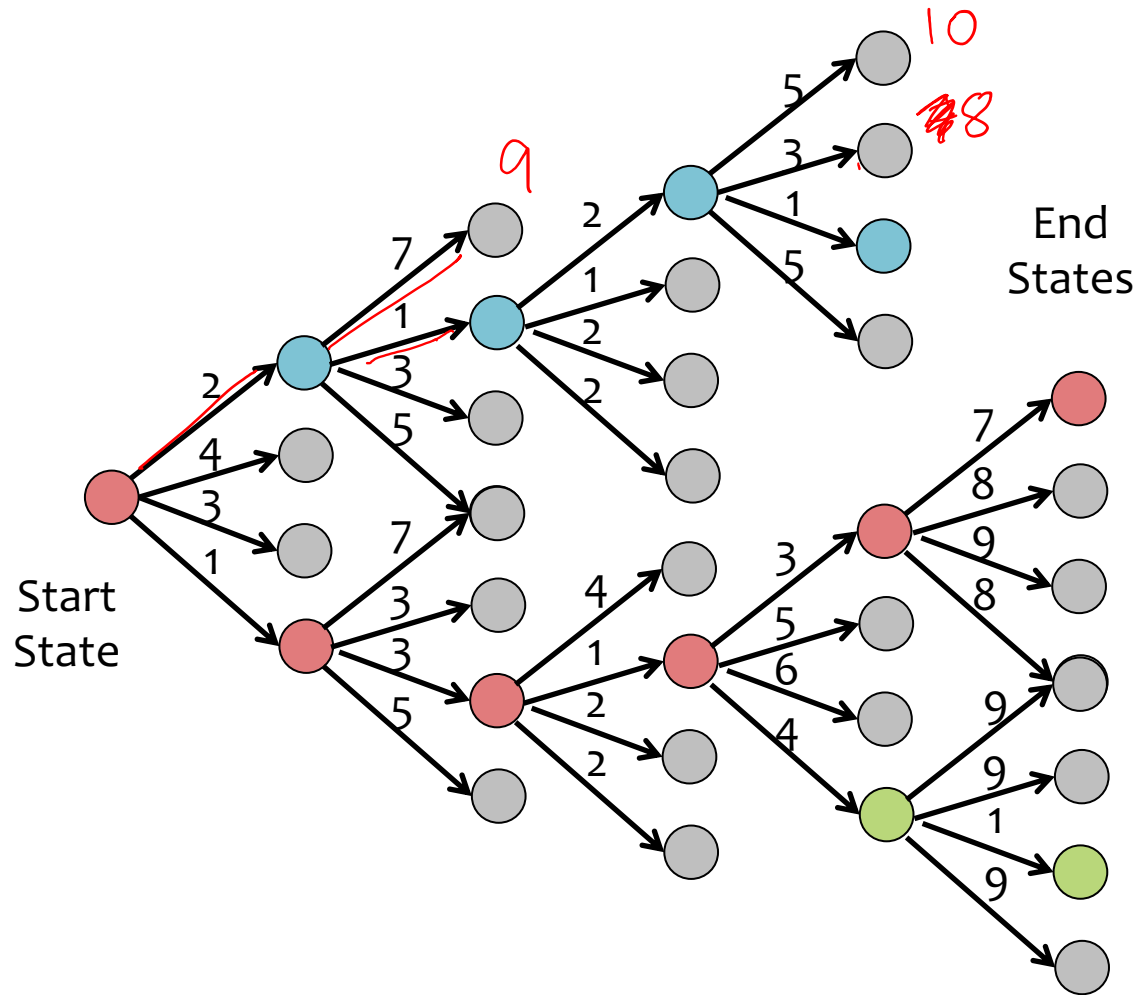
Goal:

- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:

- At each node, selects the edge with lowest (immediate) weight
- **Heuristic** method of search (i.e. does not necessarily find the best path)
- Computation time: **linear** in max path length

Background: Global Search



Goal:

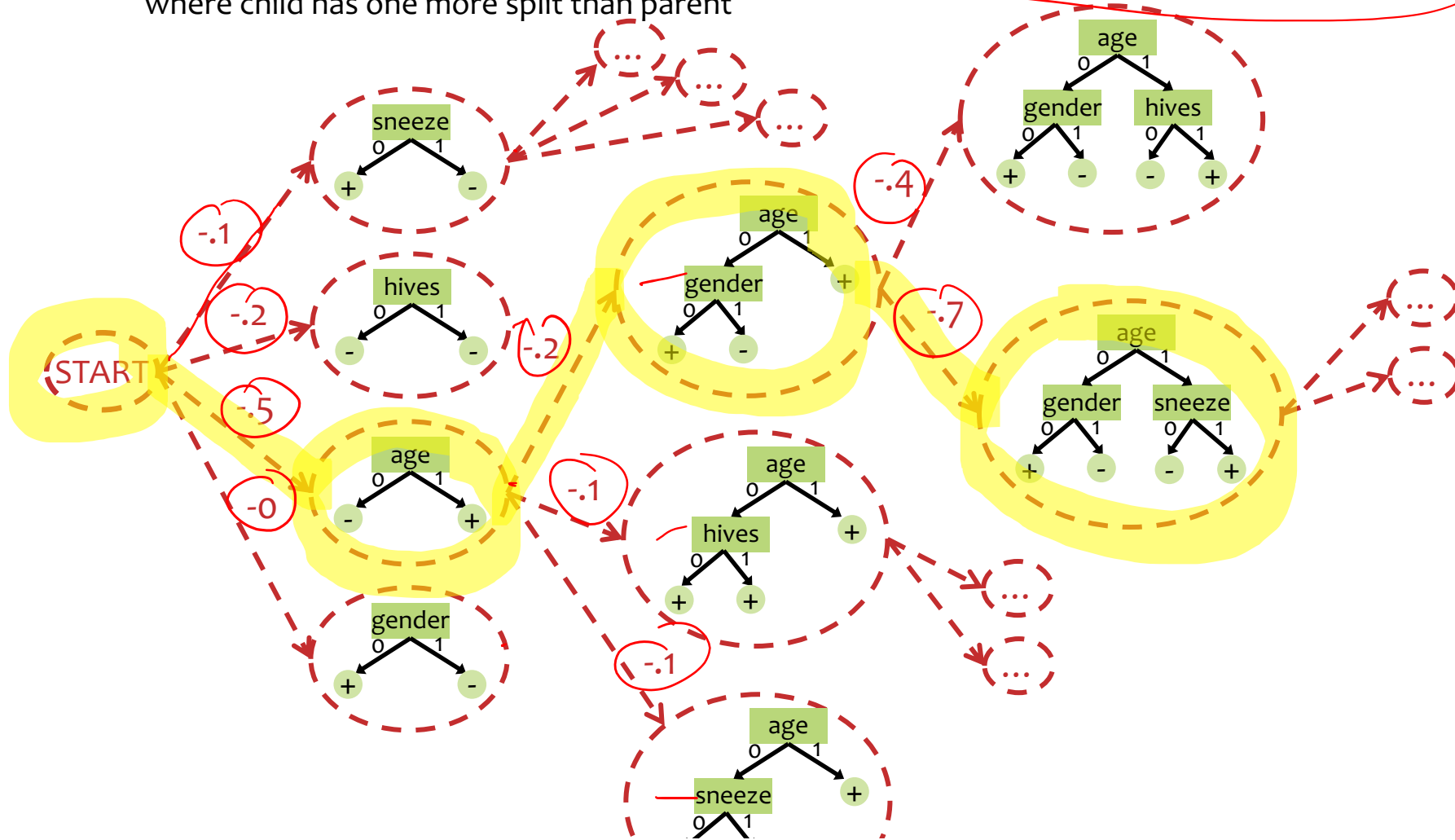
- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Global Search:

- Compute the weight of the path to **every** leaf
- **Exact** method of search (i.e. guaranteed to find the best path)
- Computation time: **exponential** in max path length

Decision Tree Learning as Search

1. **search space:** all possible decision trees
2. **node:** single decision tree
3. **edge:** connects one full tree to another, where child has one more split than parent
4. **edge weight:** ~~(negative) splitting criterion~~
5. **DT learning:** greedy search, maximizing our splitting criterion at each step



Big Question:

How is it that your
ML algorithm can
generalize to
unseen examples?

DT: Remarks

ID₃ = Decision Tree
Learning with Mutual
Information as the
splitting criterion

Question: Which tree does ID₃ find?

Definition:

We say that the **inductive bias** of a machine learning algorithm is the principal by which it generalizes to unseen examples

What is the inductive bias of ID₃?

DT: Remarks

ID₃ = Decision Tree Learning with Mutual Information as the splitting criterion

Question: Which tree does ID₃ find?

Definition:

We say that the **inductive bias** of a machine learning algorithm is the principal by which it generalizes to unseen examples

What is the inductive bias of ID₃?

Greedy search for the smallest tree that matches the data with high mutual information attributes near the top

Occam's Razor: (restated for ML)

Prefer the simplest hypothesis that explains the data

Decision Tree Learning Example

Dataset:

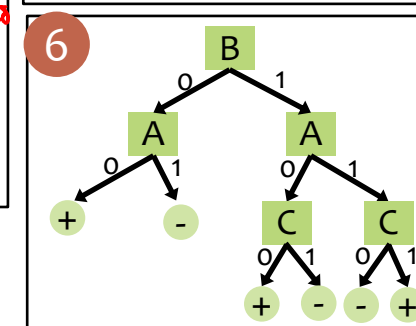
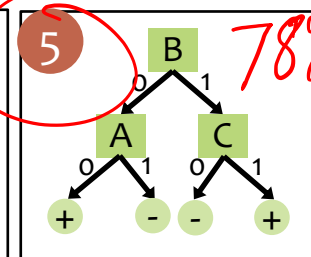
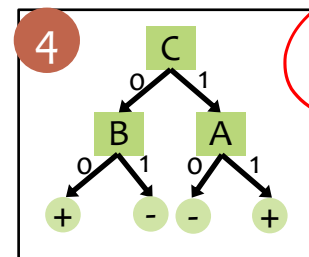
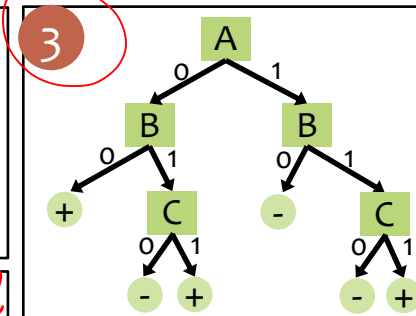
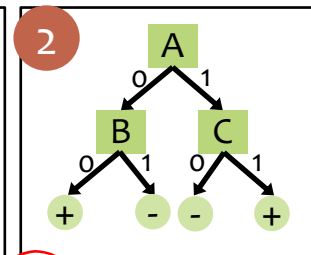
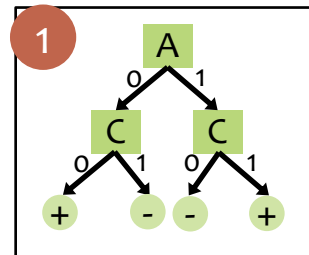
Output Y, Attributes A, B, C

Y	A	B	C
+	0	0	0
+	0	0	1
-	0	1	0
+	0	1	1
-	1	0	0
-	1	0	1
-	1	1	0
+	1	1	1

In-Class Exercise

Suppose you had an algorithm that found **the tree with lowest training error that was as small as possible (i.e. exhaustive global search)**, which tree would it return?

(Assume ties are broken by choosing the smallest.)



OVERFITTING (FOR DECISION TREES)

Decision Tree Generalization

Question:

Which of the following would generalize best to unseen examples?

- A. Small tree with low training accuracy
- ~~B. Large tree with low training accuracy~~
- 81% C. Small tree with high training accuracy
- D. Large tree with high training accuracy

Answer:

Overfitting and Underfitting

Underfitting

- The model...
 - is too simple
 - is unable captures the trends in the data
 - exhibits too much bias
- *Example:* majority-vote classifier (i.e. depth-zero decision tree)
- *Example:* a toddler (that has **not** attended medical school) attempting to carry out medical diagnosis

Overfitting

- The model...
 - is too complex
 - is fitting the noise in the data or fitting “outliers”
 - does not have enough bias
- *Example:* our “memorizer” algorithm responding to an irrelevant attribute
- *Example:* medical student who simply memorizes patient case studies, but does not understand how to apply knowledge to new patients

Overfitting

- Given a hypothesis h , its...

... error rate over all training data: $\text{error}(h, D_{\text{train}})$

... error rate over all test data: $\text{error}(h, D_{\text{test}})$

... true error over all data: $\text{error}_{\text{true}}(h)$

- We say h overfits the training data if...

$$\text{error}_{\text{true}}(h) > \text{error}(h, D_{\text{train}})$$

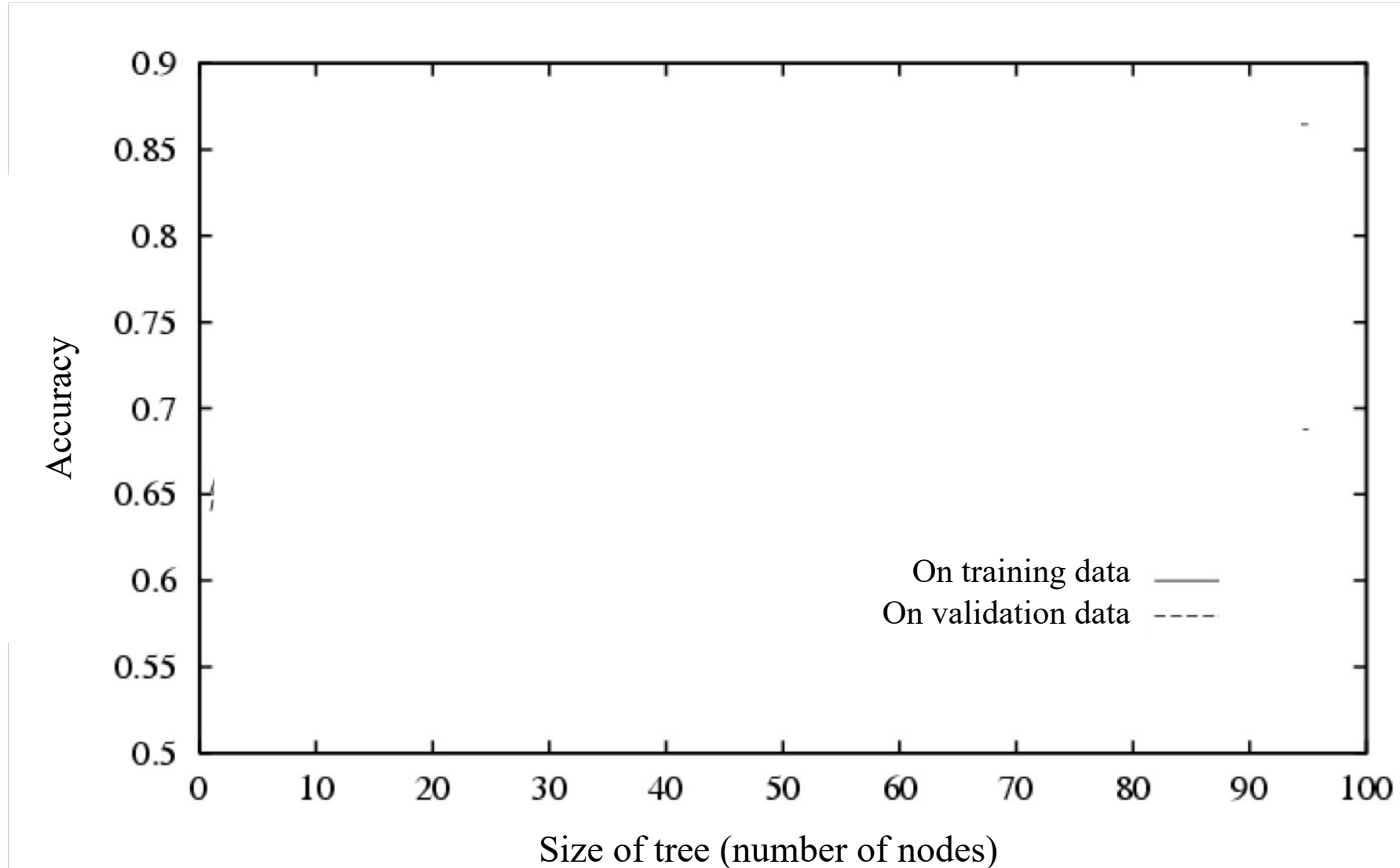
- Amount of overfitting =

$$\text{error}_{\text{true}}(h) - \text{error}(h, D_{\text{train}})$$



In practice,
 $\text{error}_{\text{true}}(h)$ is
unknown

Overfitting in Decision Tree Learning



1 - error

Figure from Tom Mitchell

Overfitting in Decision Tree Learning

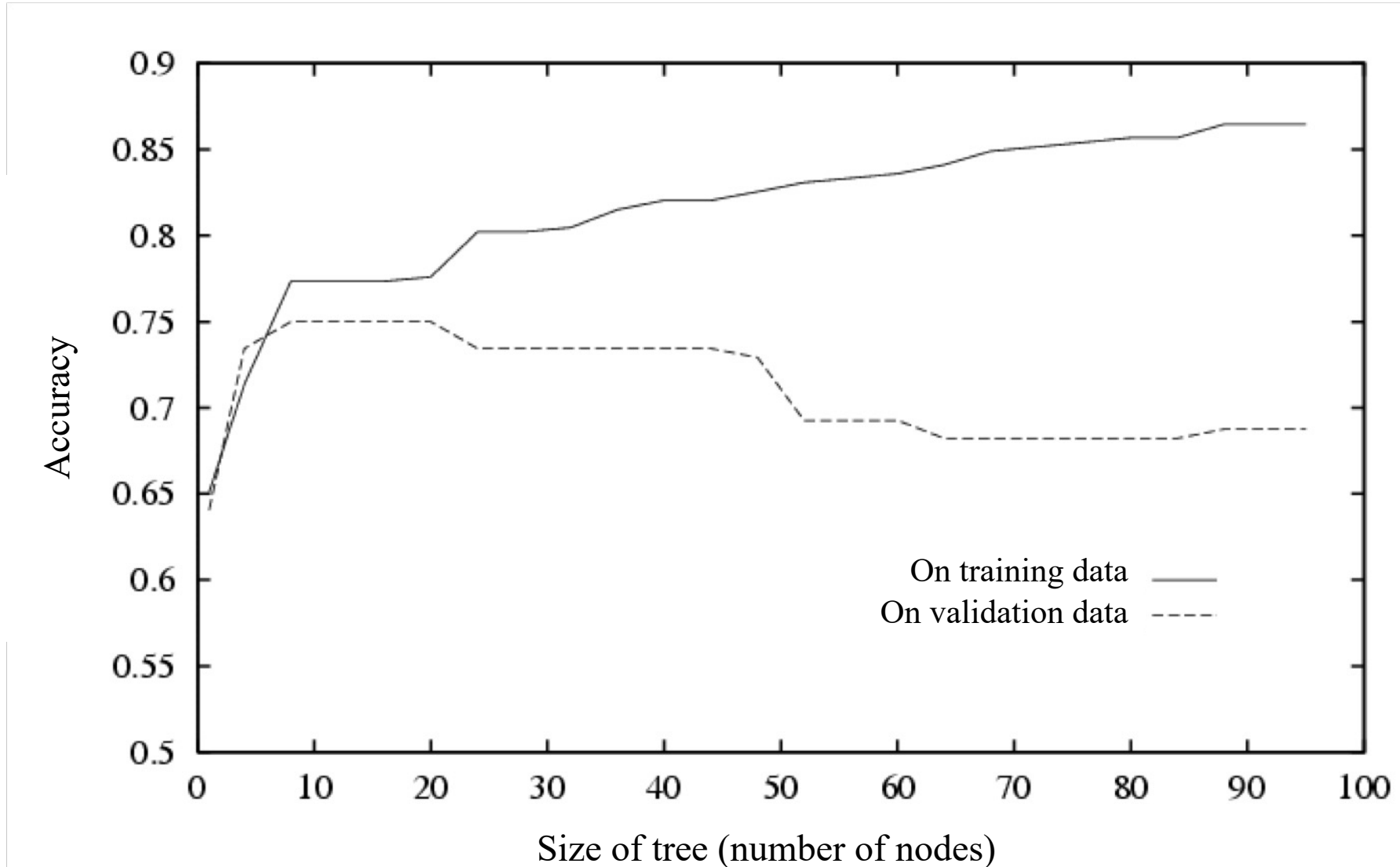


Figure from Tom Mitchell

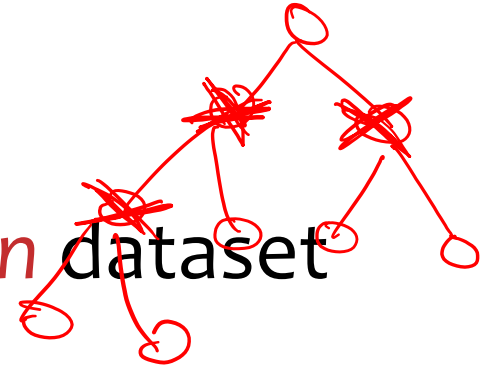
How to Avoid Overfitting?

For Decision Trees...

1. Do not grow tree beyond some **maximum depth**
2. Do not split if splitting criterion (e.g. mutual information) is **below some threshold**
3. Stop growing when the split is **not statistically significant**
4. Grow the entire tree, then **prune**

Reduced Error Pruning

1. Split data in two: *training* dataset and *validation* dataset
2. Grow the full tree using the *training* dataset
3. Repeatedly prune the tree:
 - Evaluate each split using a *validation* dataset by comparing the validation error rate **with and without** that split
 - (Greedy) remove the split that most decreases the validation error rate
 - Stop if no split improves validation error, otherwise repeat



Reduced Error Pruning

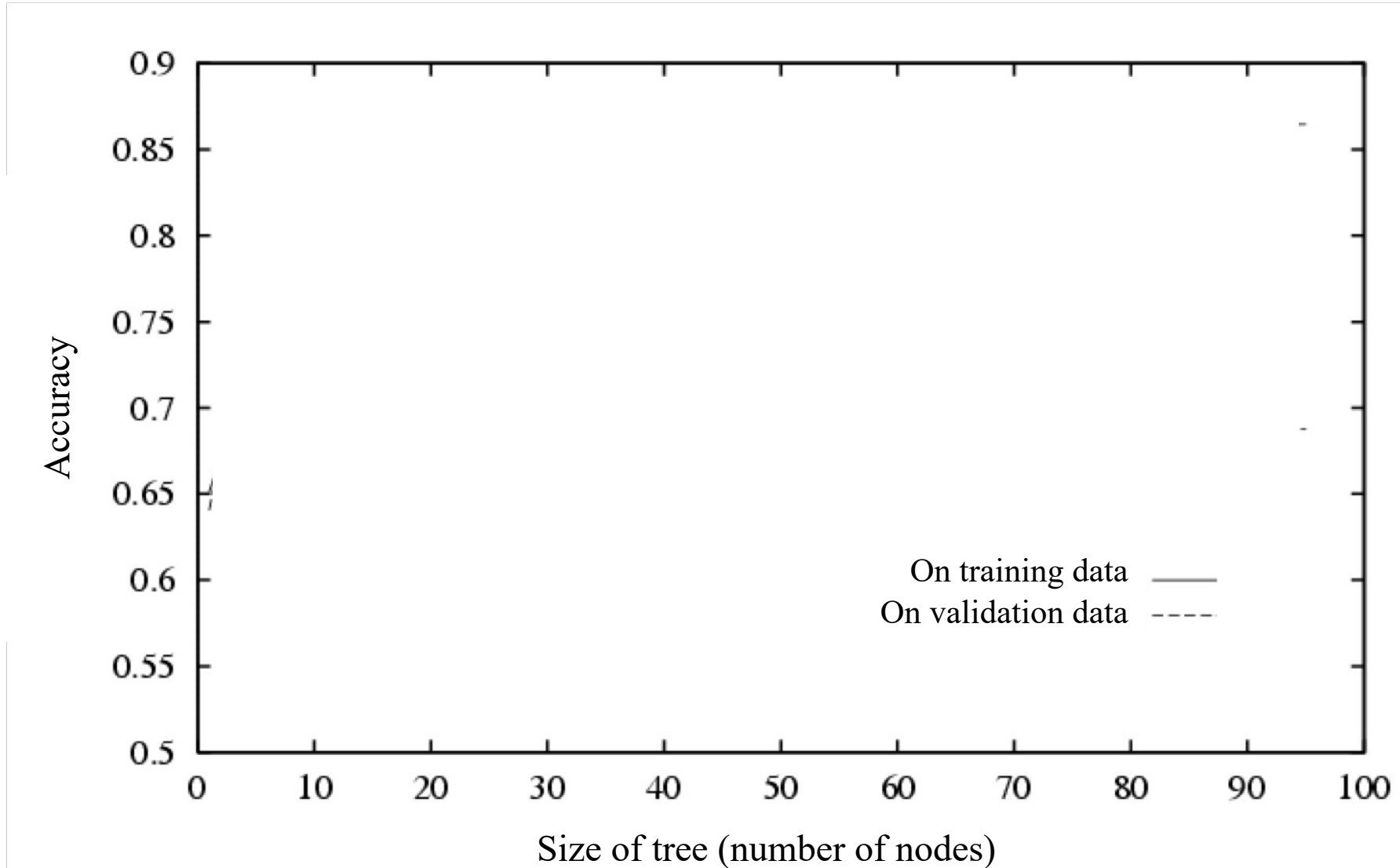


Figure from Tom Mitchell

Reduced Error Pruning

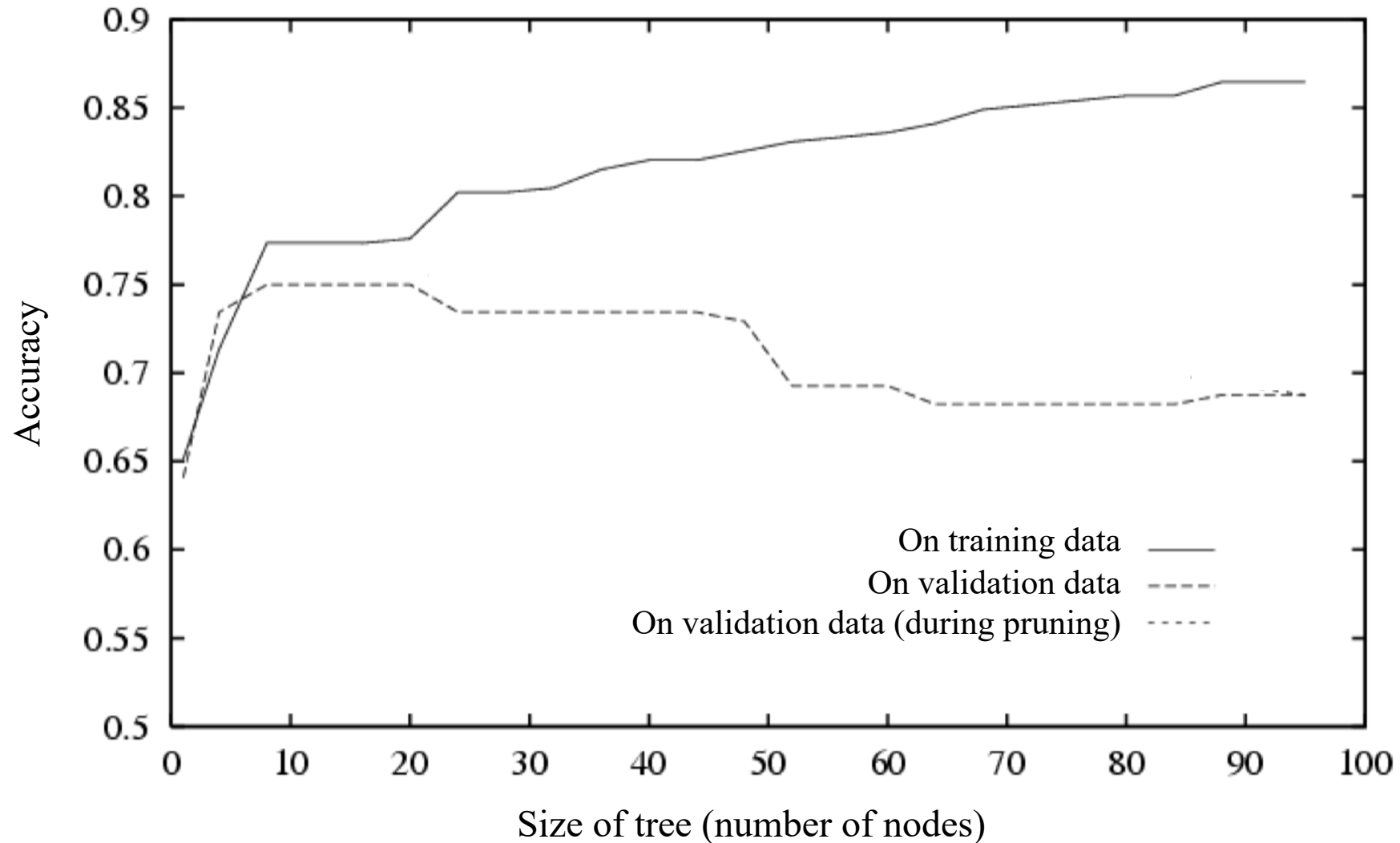
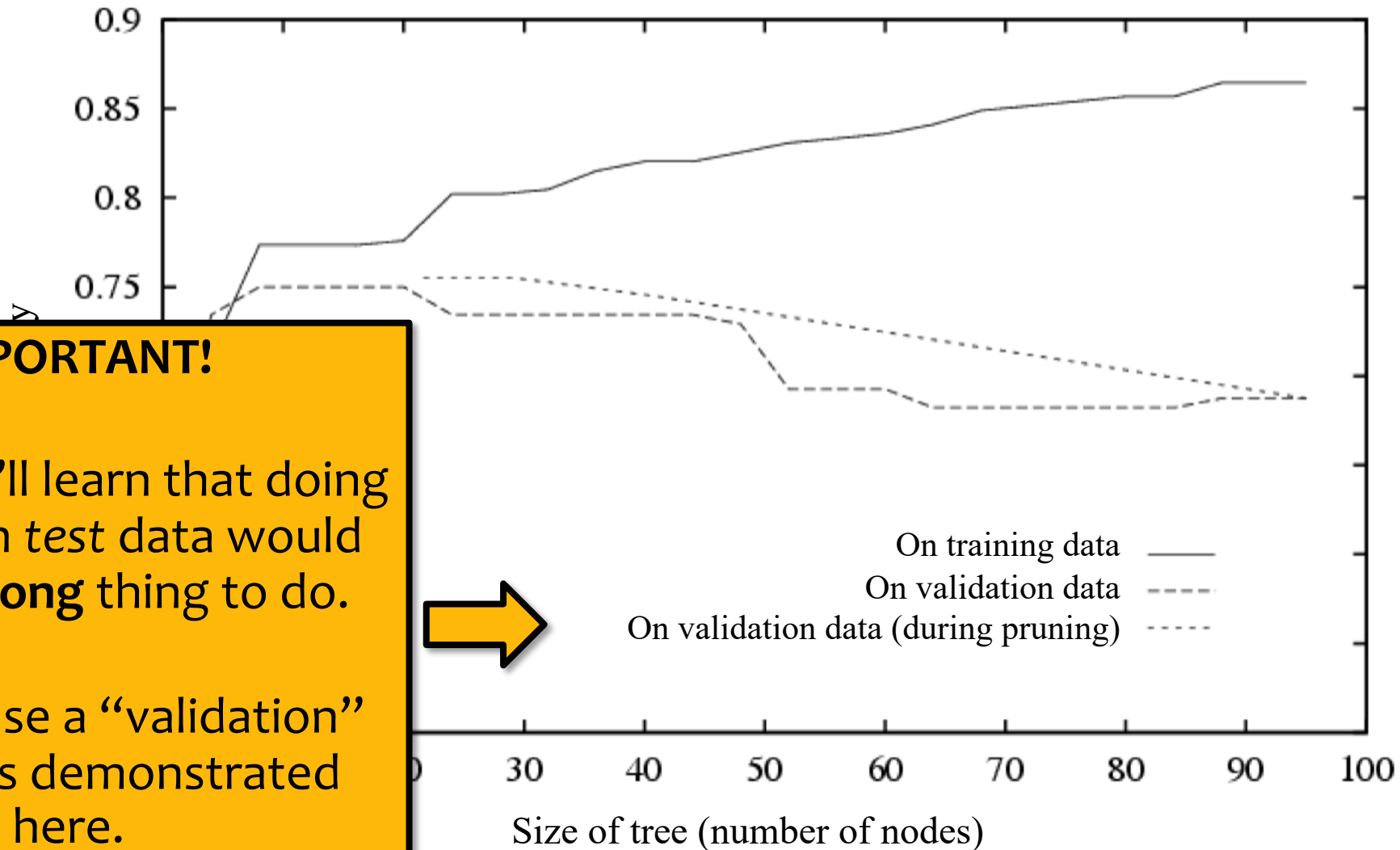


Figure from Tom Mitchell

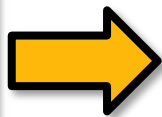
Reduced Error Pruning



IMPORTANT!

Shortly, we'll learn that doing pruning on *test* data would be the **wrong** thing to do.

We must use a "validation" dataset as demonstrated here.



Decision Trees (DTs) in the Wild

- DTs are one of the most popular classification methods for practical applications
 - Reason #1: The learned representation is **easy to explain** a non-ML person
 - Reason #2: They are **efficient** in both computation and memory
- DTs can be applied to a wide variety of problems including **classification, regression, density estimation**, etc.
- **Applications of DTs** include...
 - medicine, molecular biology, text classification, manufacturing, astronomy, agriculture, and many others
- **Decision Forests** learn many DTs from random subsets of features; the result is a very powerful example of an **ensemble method** (discussed later in the course)

DT Learning Objectives

You should be able to...

1. Implement Decision Tree training and prediction
2. Use effective splitting criteria for Decision Trees and be able to define entropy, conditional entropy, and mutual information / information gain
3. Explain the difference between memorization and generalization [CIML]
4. Describe the inductive bias of a decision tree
5. Formalize a learning problem by identifying the input space, output space, hypothesis space, and target function
6. Explain the difference between true error and training error
7. Judge whether a decision tree is "underfitting" or "overfitting"
8. Implement a pruning or early stopping method to combat overfitting in Decision Tree learning

REAL VALUED ATTRIBUTES



Fisher Iris Dataset

Fisher (1936) used 150 measurements of flowers from 3 different species: Iris setosa (0), Iris virginica (1), Iris versicolor (2) collected by Anderson (1936)

Species	Sepal Length	Sepal Width	Petal Length	Petal Width
0	4.3	3.0	1.1	0.1
0	4.9	3.6	1.4	0.1
0	5.3	3.7	1.5	0.2
1	4.9	2.4	3.3	1.0
1	5.7	2.8	4.1	1.3
1	6.3	3.3	4.7	1.6
1	6.7	3.0	5.0	1.7

Fisher Iris Dataset

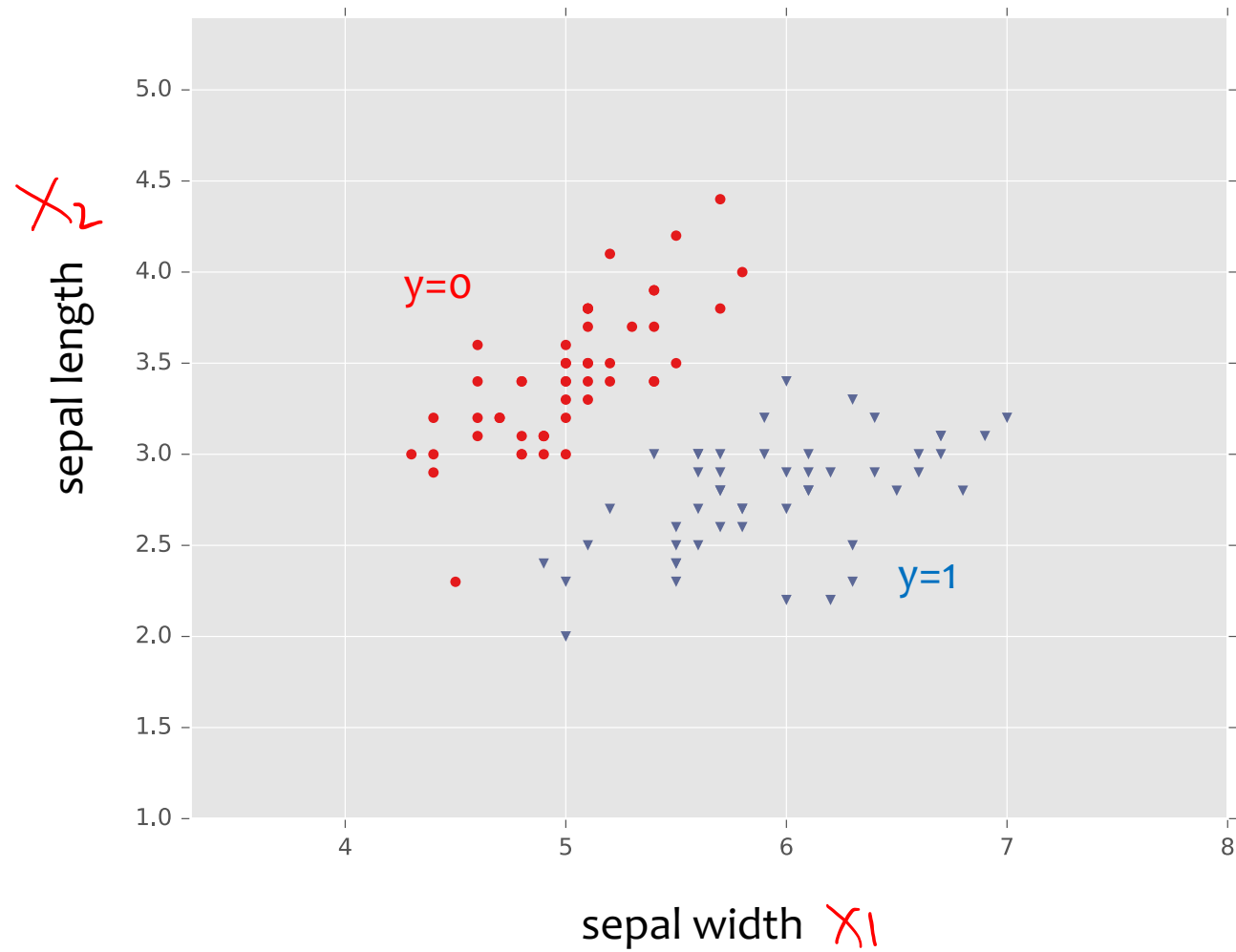
Fisher (1936) used 150 measurements of flowers from 3 different species: Iris setosa (0), Iris virginica (1), Iris versicolor (2) collected by Anderson (1936)

Species	Sepal Length	Sepal Width
0	4.3	3.0
0	4.9	3.6
0	5.3	3.7
1	4.9	2.4
1	5.7	2.8
1	6.3	3.3
1	6.7	3.0

Deleted two of the four features, so that input space is 2D



Fisher Iris Dataset





K-NEAREST NEIGHBORS

Nearest Neighbor: Algorithm

def train(\mathcal{D}):

 Store \mathcal{D}

def h(x'):

 Let $x^{(i)}$ = the point in \mathcal{D} that is nearest to x'

return $y^{(i)}$

Classification & Real-Valued Features

Classification

$$D = \{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^N$$

$\forall i \quad \vec{x}^{(i)} \in \mathbb{R}^M \leftarrow \text{features}$

$\forall i \quad y^{(i)} \in \{1, \dots, L\} \leftarrow \text{labels}$

x

y

$N = \# \text{ training examples} = |D|$
 $M = \# \text{ features}$

Binary Classification

classification where $|y| = 2$

$$\forall i \quad y^{(i)} \in \{+, -\}$$

$\in \{\text{red, blue}\}$

$\in \{\text{cat, dog}\}$

Classification & Real-Valued Features

Decision Rules / Decision Boundaries

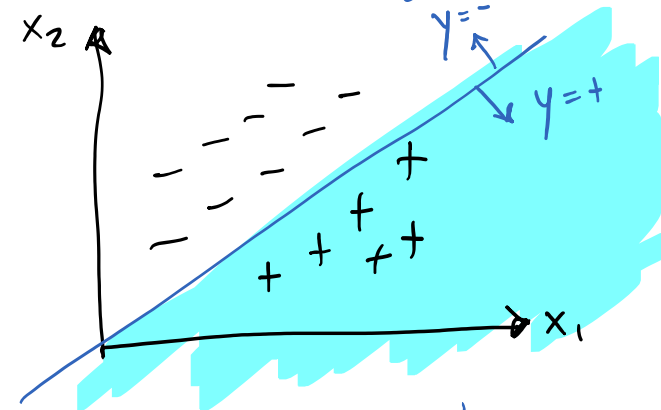
Def: decision rule / hypothesis

$$h: \mathbb{R}^M \rightarrow \{+, -\}$$

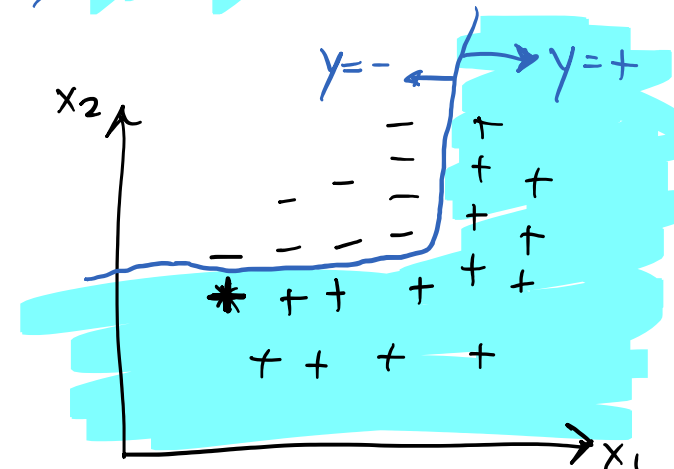
train time learn h

test time given \vec{x} , predict $\hat{y} = h(\vec{x})$

Ex: Decision Boundaries
(2D Binary Classification)
 $M=2$ $|y|=2$



linear
D. B.



nonlinear
D. B.

Nearest Neighbor: Algorithm

def train(\mathcal{D}):

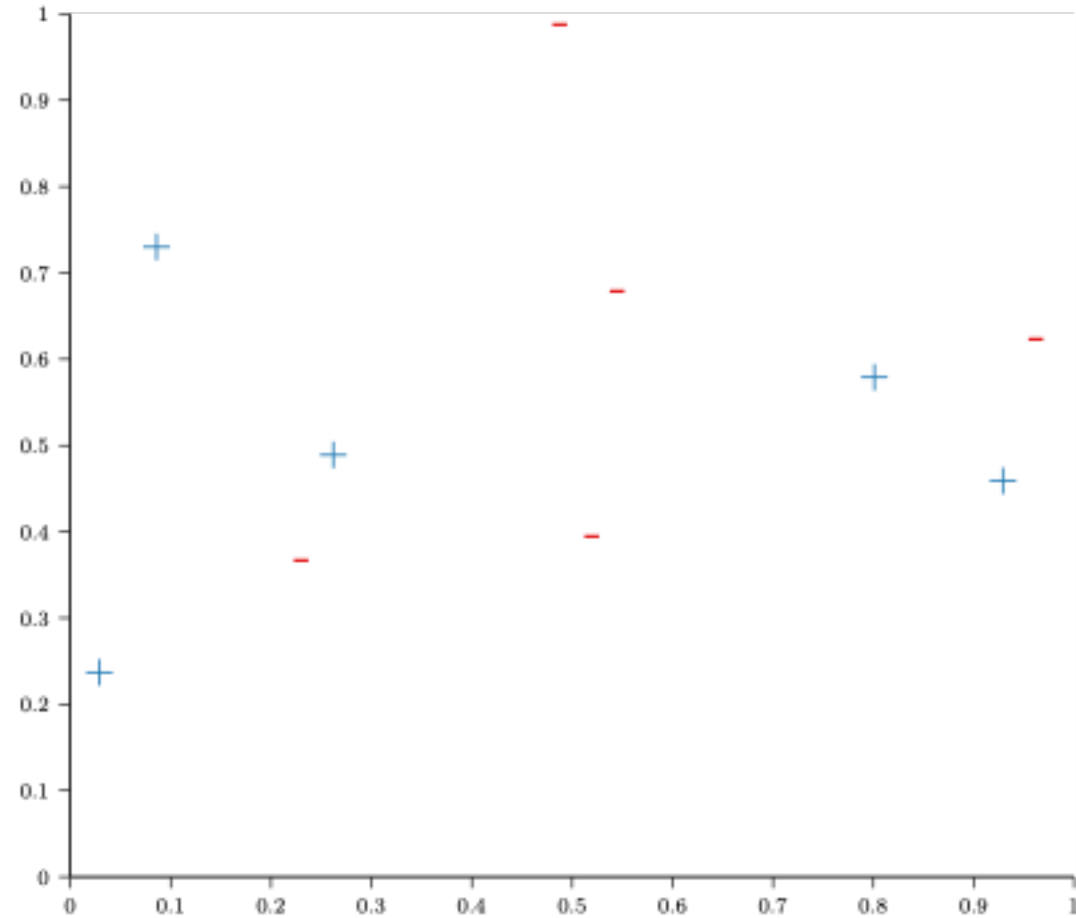
 Store \mathcal{D}

def h(x'):

 Let $x^{(i)}$ = the point in \mathcal{D} that is nearest to x'

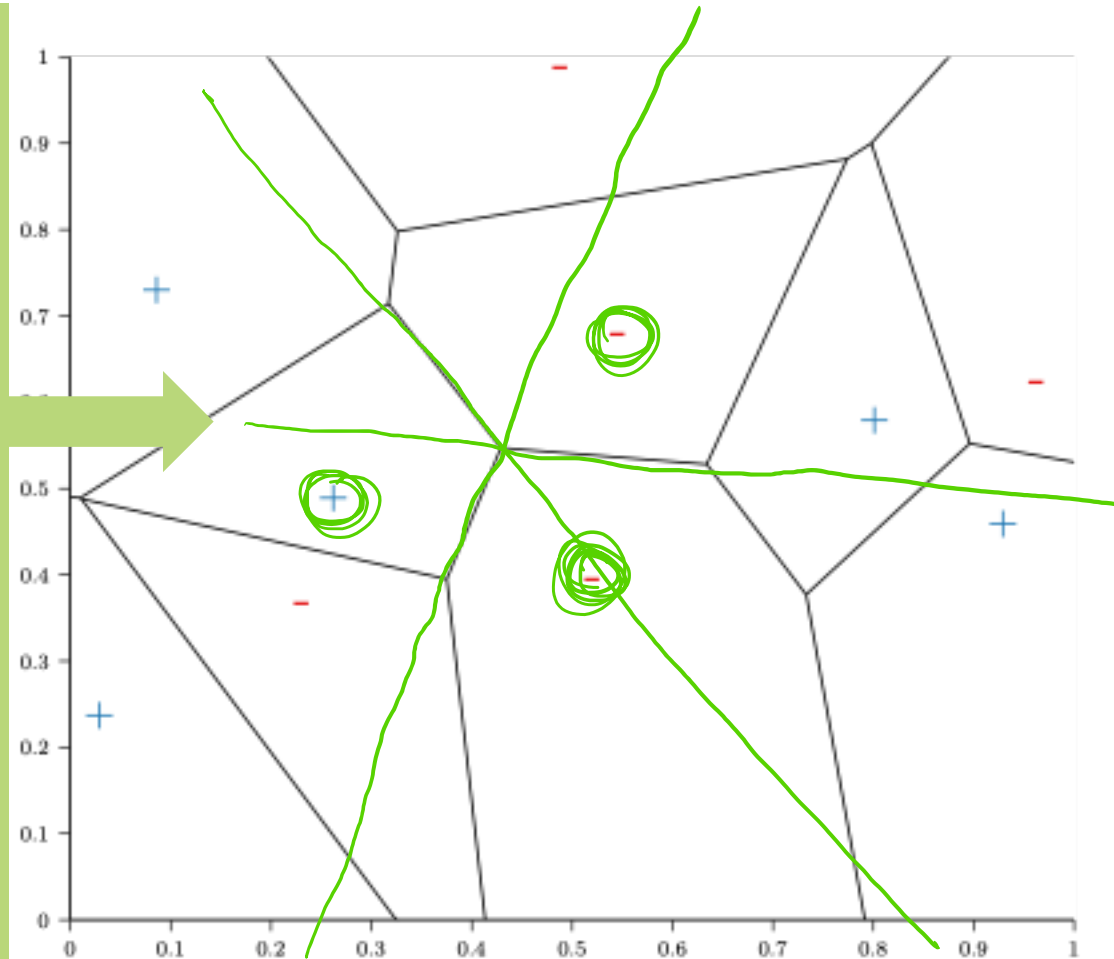
return $y^{(i)}$

Nearest Neighbor: Example

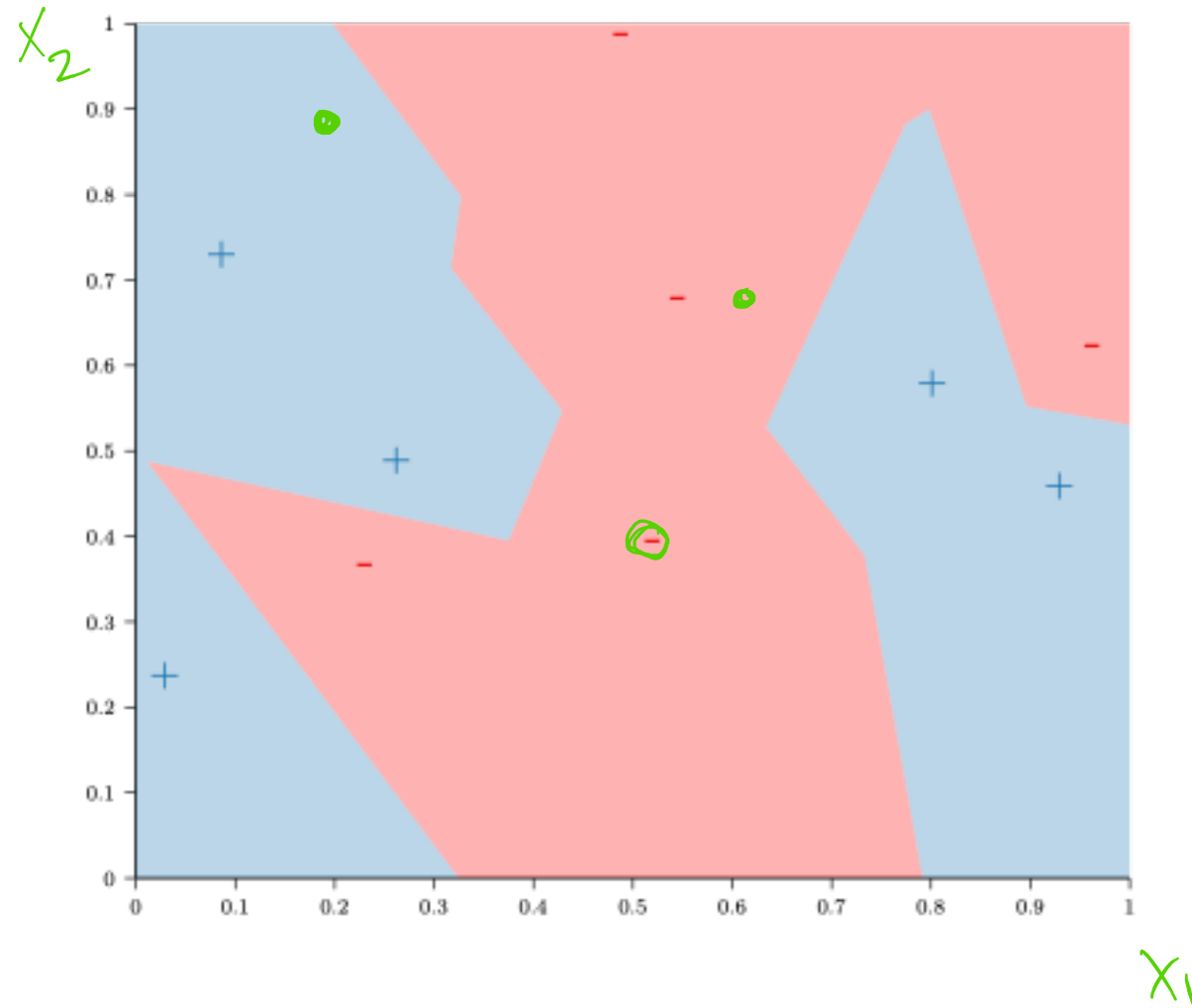


Nearest Neighbor: Example

- This is a **Voronoi diagram**
- Each **cell** contain one of our training examples
- **All points within a cell** are **closer** to that training example, than to any other training example
- **Points on the Voronoi line segments** are **equidistant** to one or more training examples



Nearest Neighbor: Example



The Nearest Neighbor Model

- Requires no training!
- Always has zero training error!
 - *A data point is always its own nearest neighbor*

k-Nearest Neighbors: Algorithm

```
def set_hyperparameters(k, d):
```

```
    Store k
```

```
    Store  $d(\cdot, \cdot)$ 
```

```
def train( $\mathcal{D}$ ):
```

```
    Store  $\mathcal{D}$ 
```

```
def h( $x'$ ):
```

```
    Let  $S$  = the set of  $k$  points in  $\mathcal{D}$  nearest to  $x'$   
    according to distance function
```

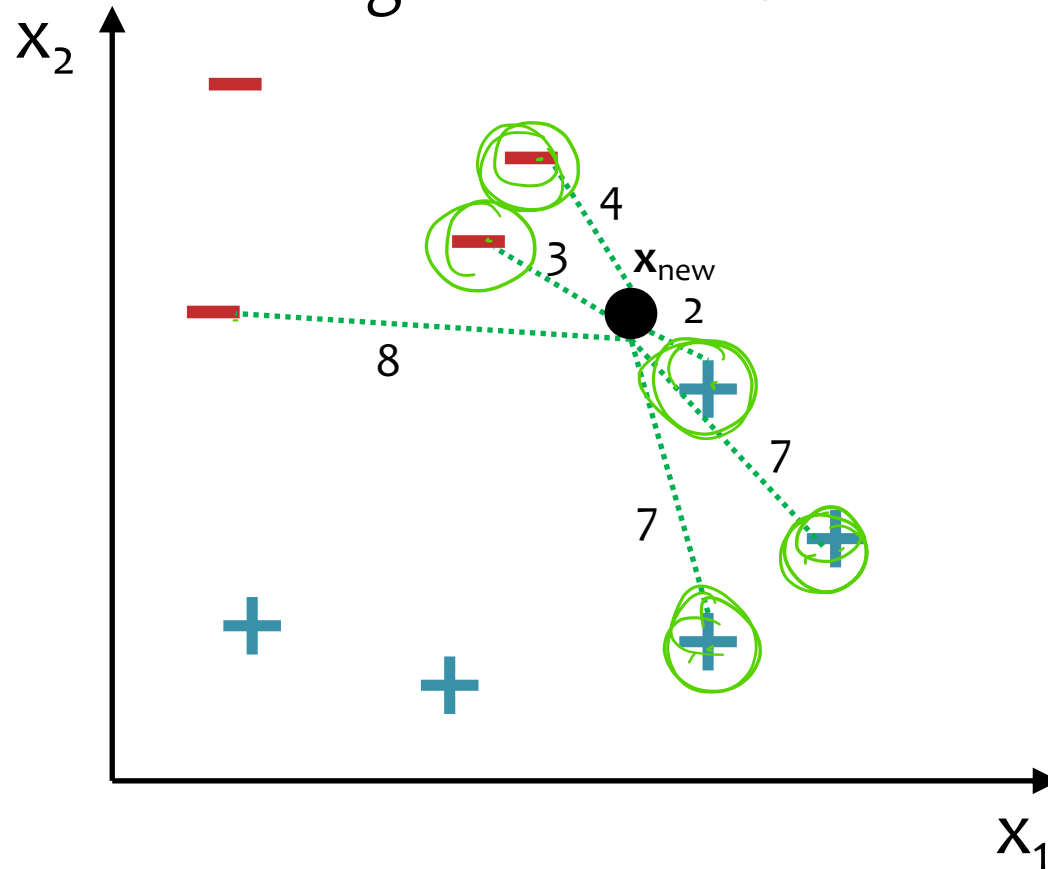
```
     $d(\mathbf{u}, \mathbf{v})$ 
```

```
    Let  $v$  = majority_vote( $S$ )
```

```
    return  $v$ 
```

k-Nearest Neighbors

Suppose we have the training dataset below.



How should we label the new point?

It depends on k:

if $k=1$, $h(\mathbf{x}_{\text{new}}) = +1$

if $k=3$, $h(\mathbf{x}_{\text{new}}) = -1$

if $k=5$, $h(\mathbf{x}_{\text{new}}) = +1$

+



-



Fred Rogers

🌐 66 languages ▾

Article [Talk](#)

[Read](#) [View source](#) [View history](#)

From Wikipedia, the free encyclopedia



"Mister Rogers" redirects here. For the television series, see [Mister Rogers' Neighborhood](#). For the asteroid, see [26858 Misterrogers](#). For other people, see [Frederick Rogers and Rogers \(surname\)](#).

Fred McFeely Rogers (March 20, 1928 – February 27, 2003) was an American television host, author, producer, and [Presbyterian minister](#).^[1] He was the creator, showrunner, and host of the preschool television series *Mister Rogers' Neighborhood*, which ran from 1968 to 2001.

Born in [Latrobe, Pennsylvania](#), near [Pittsburgh](#), Rogers earned a bachelor's degree in music from [Rollins College](#) in 1951. He began his television career at [NBC](#) in New York, returning to Pittsburgh in 1953 to work for children's programming at NET (later [PBS](#)) television station [WQED](#). He graduated from [Pittsburgh Theological Seminary](#) with a bachelor's degree in divinity in 1962 and became a Presbyterian minister in 1963. He attended the [University of Pittsburgh's](#) Graduate School of Child Development, where he began his 30-year collaboration with child psychologist [Margaret McFarland](#). He also helped develop the children's shows *The Children's Corner* (1955) for [WQED](#) in [Pittsburgh](#) and *Misterogers* (1963) in [Canada](#) for the [Canadian Broadcasting Corporation](#). In 1968, he returned to Pittsburgh and adapted the format of his Canadian series to create *Mister Rogers' Neighborhood*. It ran for 33 years and was critically acclaimed for focusing on children's emotional and physical concerns, such as death, sibling rivalry, school enrollment, and divorce.

Rogers died of stomach cancer in 2003, aged 74. His work in children's television has been widely lauded, and he received more than 40 honorary degrees and several awards, including the [Presidential Medal of Freedom](#) in 2002 and a [Lifetime Achievement Emmy](#) in 1997. He was inducted into the [Television Hall of Fame](#) in 1999. Rogers influenced many writers and producers of children's television shows, and his broadcasts provided comfort during tragic events, even after his death.

Early life

Rogers was born on March 20, 1928, at 705 Main Street in [Latrobe, Pennsylvania](#), about 40 miles (64 km) outside of [Pittsburgh](#).^[2] His father, James Hillis Rogers, was "a very successful businessman"^[3] who was president of the McFeely Brick Company, one of Latrobe's most prominent businesses. His mother, Nancy (née McFeely), [knitted](#) sweaters for American soldiers from western Pennsylvania who were fighting in Europe and regularly volunteered at the Latrobe Hospital. Initially dreaming of becoming a doctor, she settled

The Reverend
Fred Rogers



Rogers in 1982

Born	<div>Fred McFeely Rogers</div> March 20, 1928 <div>Latrobe, Pennsylvania, U.S.</div>
Died	<div>February 27, 2003 (aged 74)</div> Pittsburgh, Pennsylvania , U.S.
Other names	Mister Rogers
Education	Dartmouth College Rollins College (BM) Pittsburgh Theological Seminary (BDiv)
Occupation(s)	Children's television presenter, actor, puppeteer, singer,

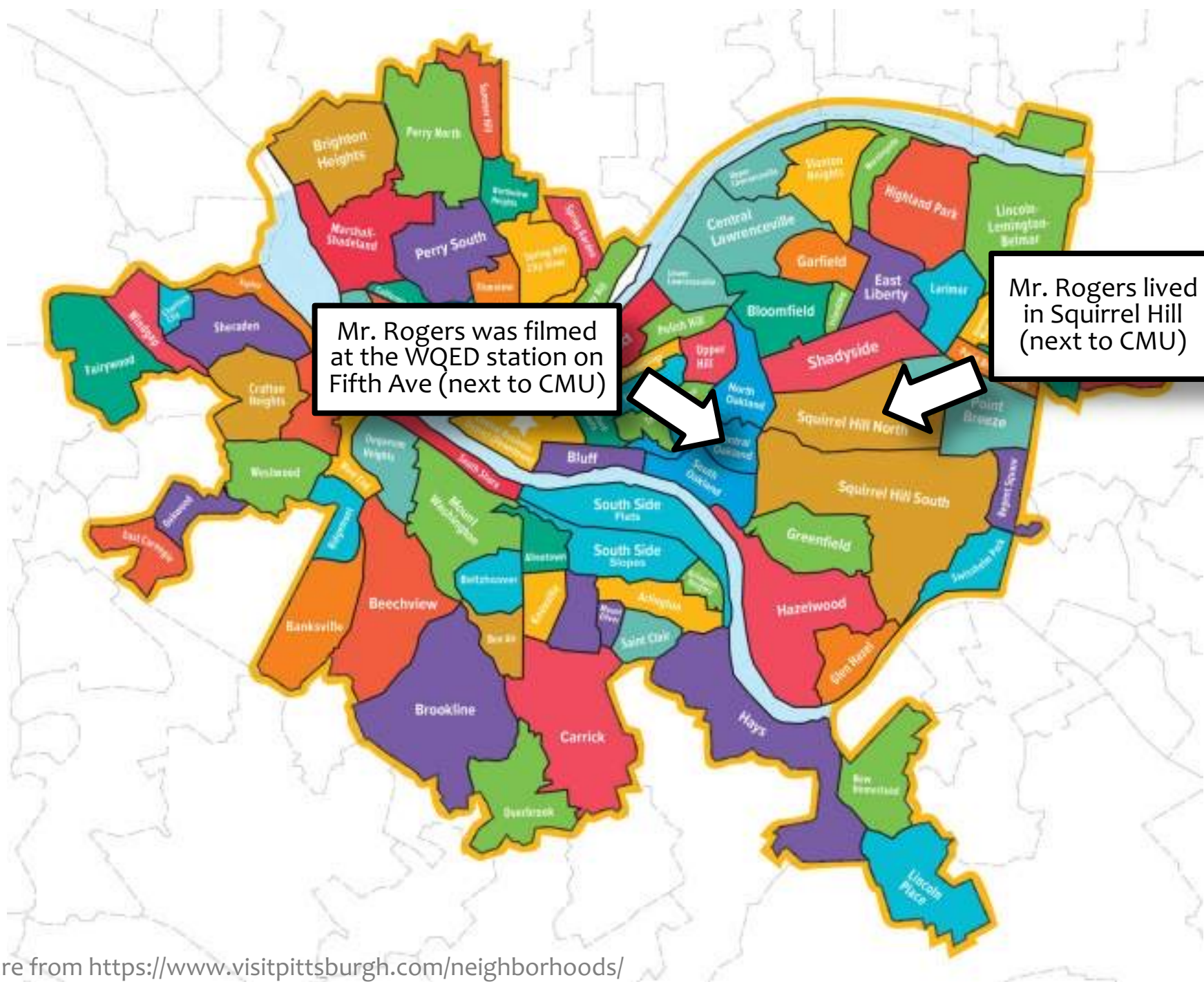


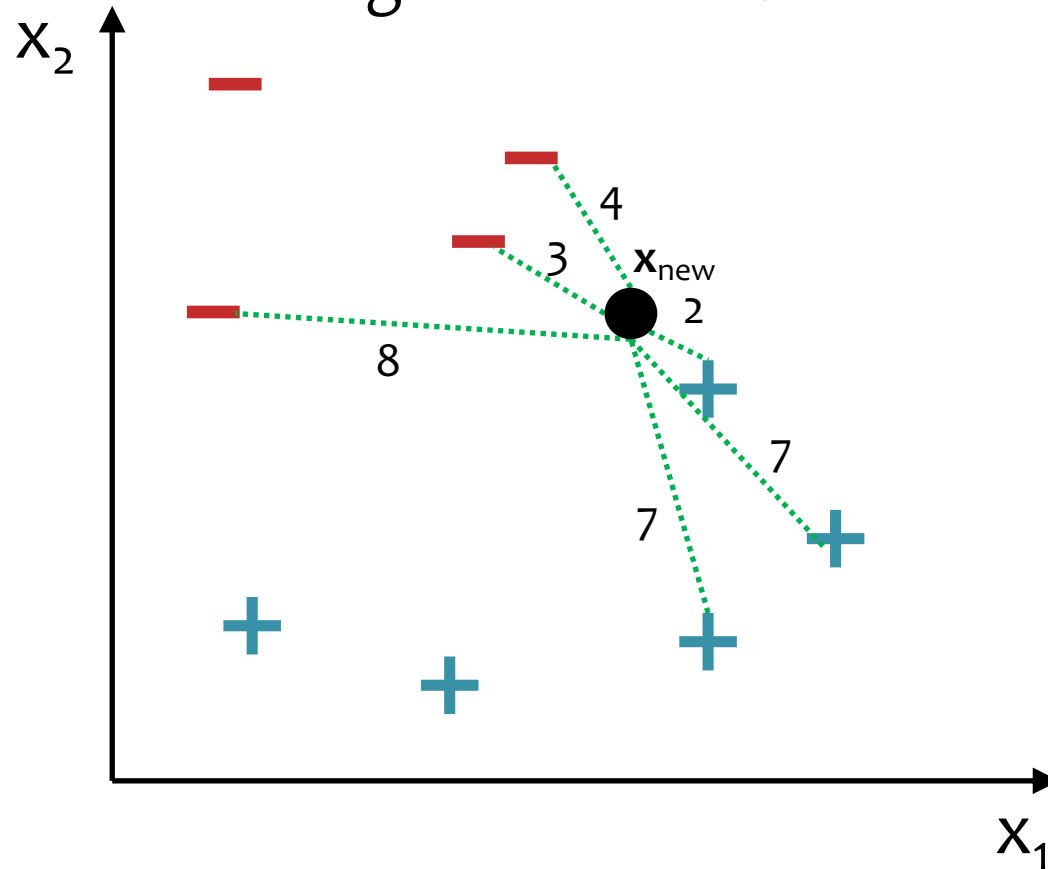
Figure from <https://www.visitpittsburgh.com/neighborhoods/>

Mr. Roger's Neighborhood

- Some of Mr. Roger's neighbors...
 - Julia Childs (cookbook author)
<https://www.misterrogers.org/videos/julia-child/>
 - Yo-yo Ma (cellist) 2:38
<https://www.misterrogers.org/videos/yo-yo-ma/>
 - Silvia Earle (marine biologist) 3:00
<https://www.misterrogers.org/videos/sylvia-earle/>
 - Wynton Marsalis (trumpet player) 4:00
<https://www.misterrogers.org/videos/wynton-marsalis/>
 - Singing Won't You Be My Neighbor
<https://misterrogers.org/videos/wont-you-be-my-neighbor/>

k-Nearest Neighbors

Suppose we have the training dataset below.



How should we label the new point?

It depends on k :

if $k=1$, $h(x_{\text{new}}) = +1$

if $k=3$, $h(x_{\text{new}}) = -1$

if $k=5$, $h(x_{\text{new}}) = +1$

+



-



KNN: Remarks

Distance Functions:

- KNN requires a **distance function**

$$d : \mathbb{R}^M \times \mathbb{R}^M \rightarrow \mathbb{R}$$

- The most common choice is **Euclidean distance**

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{m=1}^M (u_m - v_m)^2}$$

- But there are other choices (e.g. **Manhattan distance**)

$$d(\mathbf{u}, \mathbf{v}) = \sum_{m=1}^M |u_m - v_m|$$