

# CS540 Introduction to Artificial Intelligence

## Lecture 6

Young Wu

Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

July 6, 2022

# Hat Game

## Quiz

- 5 kids are wearing either green or red hats in a party: they can see every other kid's hat but not their own.
- Dad said to everyone: at least one of you is wearing green hat.
- Dad asked everyone: do you know the color of your hat?
- Everyone said no. → at least 2 Q1
- Dad asked again: do you know the color of your hat?
- Everyone said no. at least 3
- Dad asked again: do you know the color of your hat?
- ~~Some kids (at least one) said yes.~~
- No one lied. How many kids are wearing green hats?
- A: 1... B: 2... C: 3... D: 4... E: 5



# Discussion Grades

Admin

- The Discussion grades for week 1 is still computed incorrectly: will try to fix tonight.
- The past exam links are fixed, please refresh the pages.
- The list of relevant past exam questions are on the Q1, Q2, etc, pages.
- Please do NOT start the homework that are not announced yet! Especially please do not share solutions to those.



# Discussion Posts

Admin

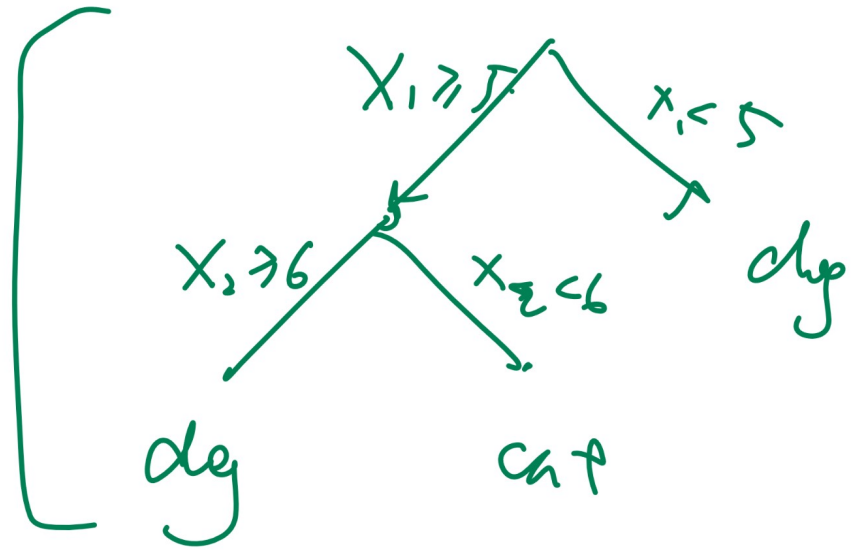
- Complete slides are listed as "Blank Slides".
- The slides used during the lectures are listed as "Blank Slides with Blank Pages for Quiz Questions".
- Shared solutions should explain how you get the solutions, a post will not get points if it just says "according to the hints, ...".
- Link to MyScript and Maple App on W1 page.

# Axes Aligned Decision Boundary

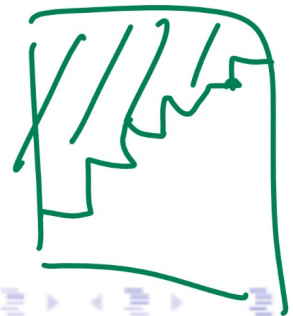
Motivation

$w^T x + b \geq 0$  ?

↳ Decision Trees



axis-aligned



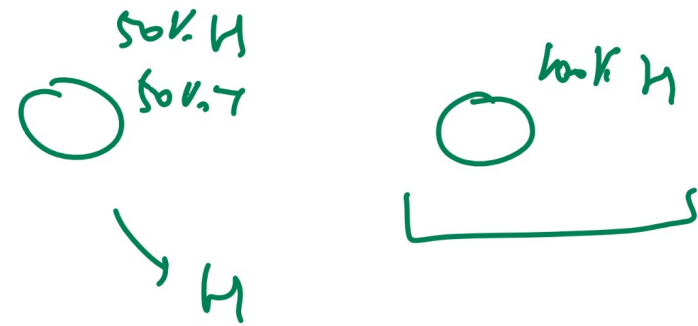
# Decision Tree

## Description

- Find the feature that is the most informative
- Split the training set into subsets according to this feature.
- Repeat on the subsets until all the labels in the subset are the same.

# Binary Entropy

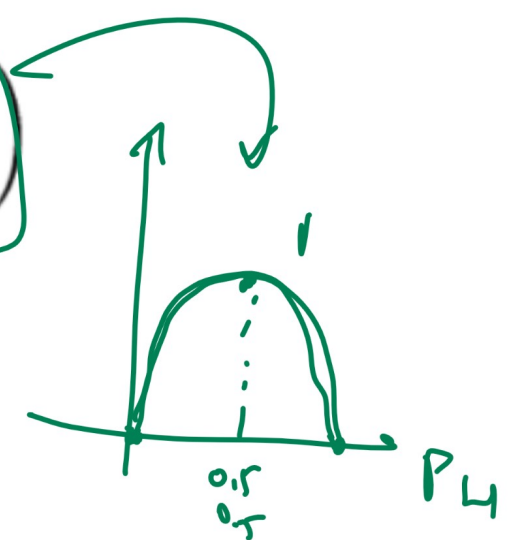
## Definition



- Entropy is the measure of uncertainty.
- The value of something uncertain is more informative than the value of something certain.
- For binary labels,  $y_i \in \{0, 1\}$ , suppose  $p_0$  fraction of labels are 0 and  $1 - p_0 = p_1$  fraction of the training set labels are 1, the entropy is:

$$H(Y) = p_0 \log_2 \left( \frac{1}{p_0} \right) + p_1 \log_2 \left( \frac{1}{p_1} \right)$$

$$= -p_0 \log_2 (p_0) - p_1 \log_2 (p_1)$$



# Entropy

## Definition

- If there are  $K$  classes and  $p_y$  fraction of the training set labels are in class  $y$ , with  $y \in \{1, 2, \dots, K\}$ , the entropy is:

$$\begin{aligned} H(Y) &= \sum_{y=1}^K p_y \log_2 \left( \frac{1}{p_y} \right) \\ &= - \sum_{y=1}^K p_y \log_2 (p_y) \end{aligned}$$

# Entropy

## Quiz

- Running from You-Know-Who, Harry enters the CS building on the 1st floor. He flips a fair coin: if it is heads he hides in room 1325; otherwise, he climbs to the 2nd floor. In that case, he flips the coin again: if it is heads he hides in CSL; otherwise, he climbs to the 3rd floor and hides in 3331. What is the entropy of Harry's location?

$$-\sum p_i \log_2 p_i$$

$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$
1325	CSL	3331

$$-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 1.5$$





# Entropy 2

## Quiz

$$\log(p_i) / \log(2)$$

Q 4

- A bag contains a red ball, a green ball, a blue ball, and a black ball. Randomly draw a ball from the bag with equal probability. What is the entropy of the outcome?

- A : 1
- B :  $\log_2(3)$
- C : 1.5
- D : 2
- E : I don't understand entropy

$$- \sum p_i \log_2 p_i$$
$$- \frac{1}{4} \log_2 \frac{1}{4}$$
$$\left( \frac{1}{2} \right) \cdot 4 = 2$$



# Conditional Entropy

## Definition

- Conditional entropy is the entropy of the conditional distribution. Let  $K_X$  be the possible values of a feature  $X$  and  $K_Y$  be the possible labels  $Y$ . Define  $p_x$  as the fraction of the instances that are  $x$ , and  $p_{y|x}$  as the fraction of the labels that are  $y$  among the ones with instance  $x$ .

$$\begin{aligned} H(Y|X=x) &= - \sum_{y=1}^{K_Y} p_{y|x} \log_2(p_{y|x}) \\ H(Y|X) &= \sum_{x=1}^{K_X} p_x H(Y|X=x) \end{aligned}$$

$\frac{c_{y|x}}{c_x}$

# Aside: Cross Entropy

## Definition

- Cross entropy measures the difference between two distributions.

$$\underline{H(Y, X)} = - \sum_{z=1}^K \underbrace{p_{Y=z}}_{\text{actual label}} \log_2 \left( \underbrace{p_{X=z}}_{\text{predicted label}} \right)$$

- It is used in logistic regression to measure the difference between actual label  $Y_i$  and the predicted label  $A_i$  for instance  $i$ , and at the same time, to make the cost convex.

$$\underline{H(Y_i, A_i) = -y_i \log(a_i) - (1 - y_i) \log(1 - a_i)}$$

# Information Gain

## Definition

- The information gain is defined as the difference between the entropy and the conditional entropy.

$$I(Y|X) = H(Y) - H(Y|X).$$

- The larger than information gain, the larger the reduction in uncertainty, and the better predictor the feature is.

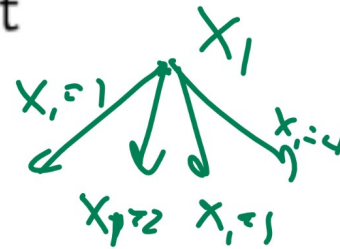
# Splitting Discrete Features

## Definition

- The most informative feature is the one with the largest information gain.

$$\operatorname{argmax}_j \underline{I(Y|X_j)}$$

1, 2, 3, 4



- Splitting means dividing the training set into  $K_{X_j}$  subsets.

$$\underline{\{(x_i, y_i) : x_{ij} = 1\}}, \{(x_i, y_i) : x_{ij} = 2\}, \dots, \{(x_i, y_i) : x_{ij} = K_{X_j}\}$$

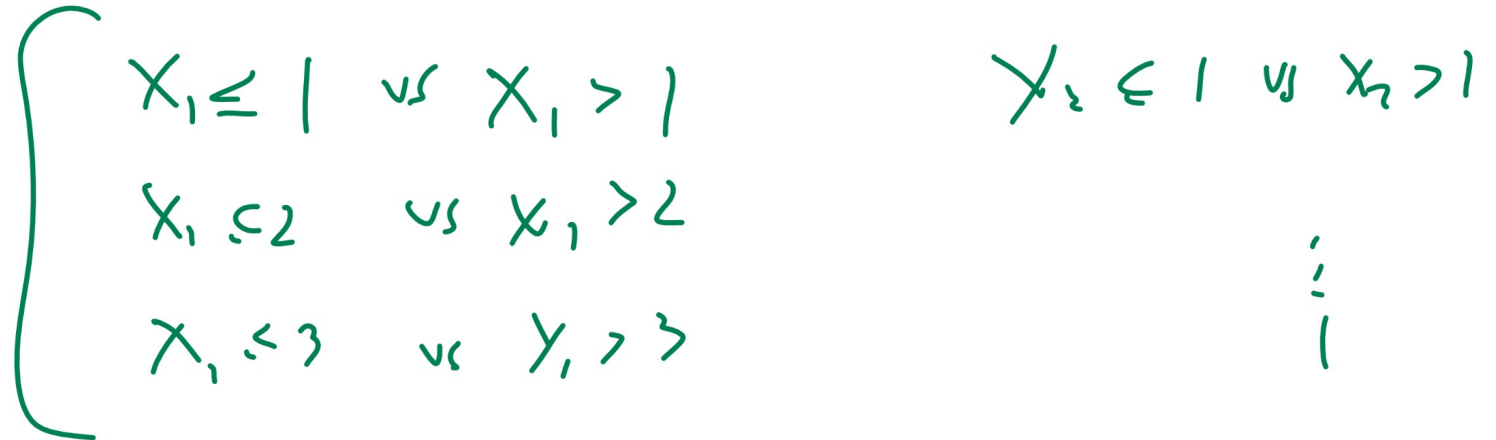
# Splitting Continuous Variables Diagram

Definition

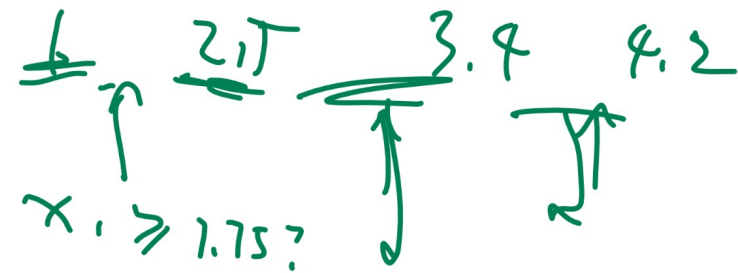


P2

$X_1$  1, 2, ..., 10



IG



# ID3 Algorithm (Iterative Dichotomiser 3)

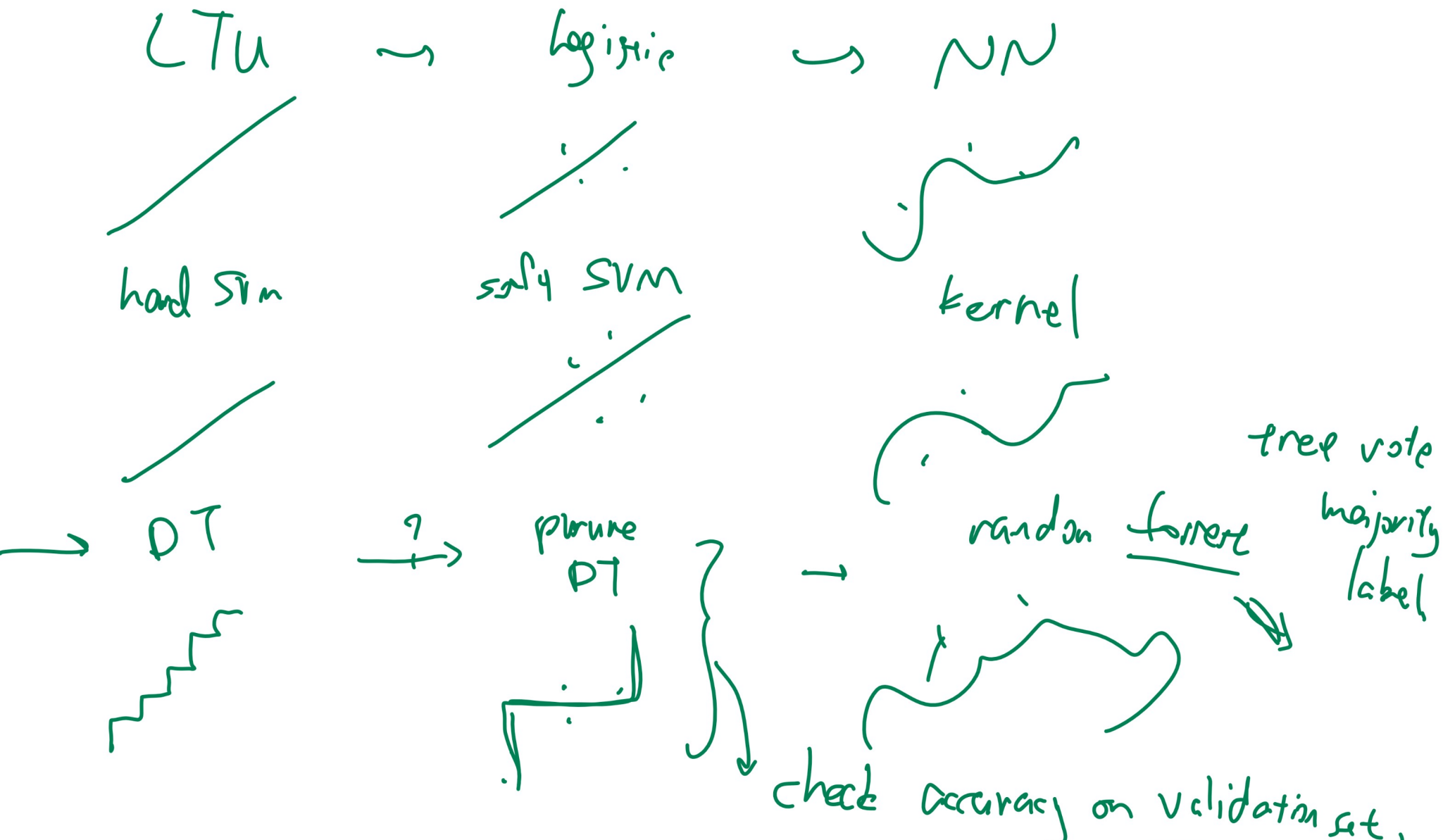
## Description

- Find the feature that is the most informative.
- Split the training set into subsets according to this feature.
- Repeat on the subsets until all the labels in the subset are the same.

Q1 → Pick a random choice  
A - E

# Pruning Diagram

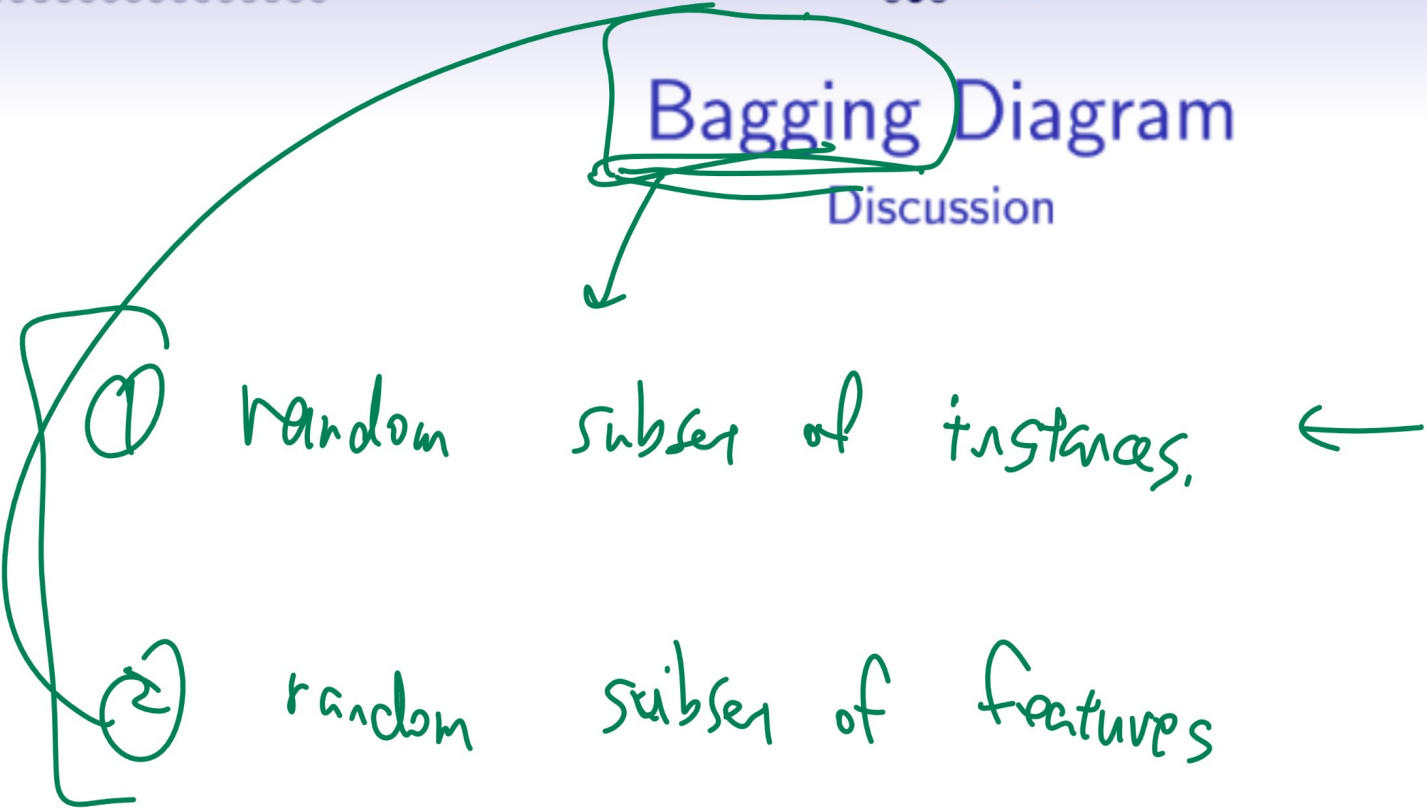
## Discussion





# Bagging Diagram

Discussion

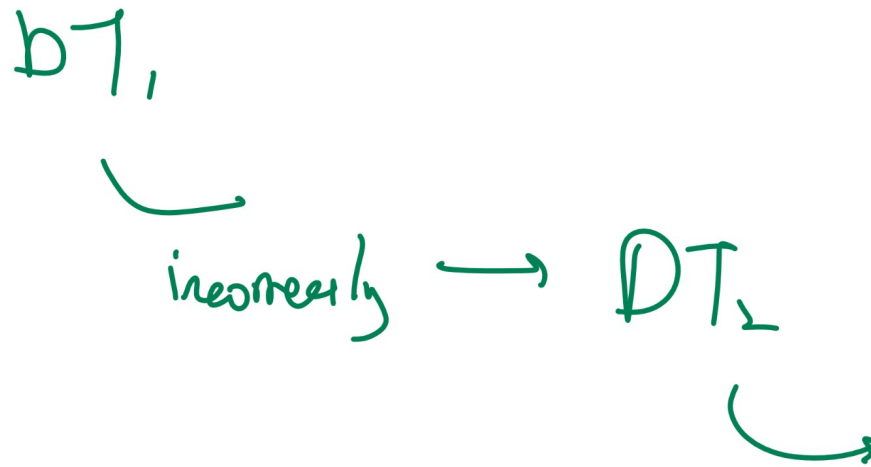


bootstrap aggregating



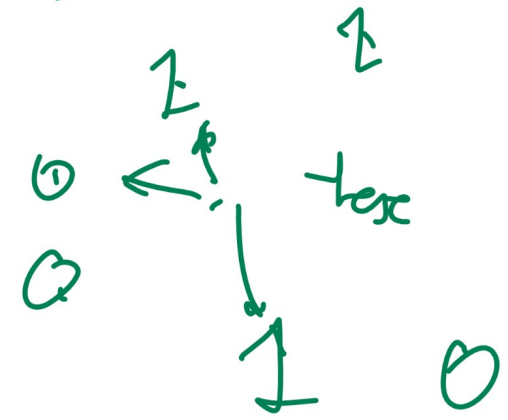
# Boosting Diagram

## Discussion



# K Nearest Neighbor 3NN

Description



- Given a new instance, find the  $K$  instances in the training set that are the closest.
- Predict the label of the new instance by the majority of the labels of the  $K$  instances.

# Distance Function

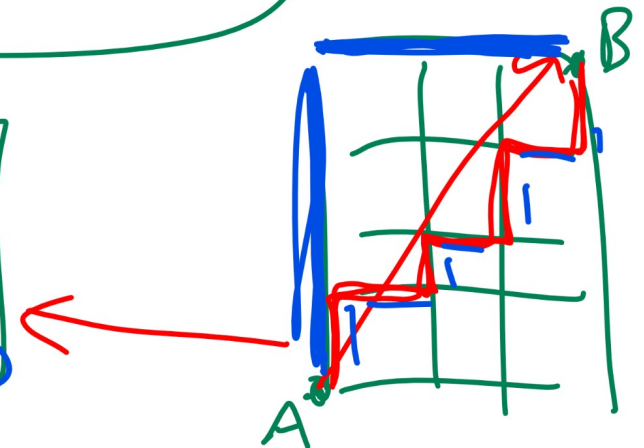
## Definition

- Many distance functions can be used in place of the Euclidean distance.

$$\rho(x, x') = \|x - x'\|_2 = \sqrt{\sum_{j=1}^m (x_j - x'_j)^2}$$

- An example is Manhattan distance.

$$\rho(x, x') = \sum_{j=1}^m |x_j - x'_j|$$





# 1 Nearest Neighbor

## Quiz

- Find the 1 Nearest Neighbor label for  $\begin{bmatrix} 3 \\ 6 \end{bmatrix}$  using Manhattan distance.

$x_1$	1	1	3	5	2
$x_2$	1	7	3	4	5
$y$	0	1	1	0	0

- A : 0
- B : 1

dist 7 3 3 4 2

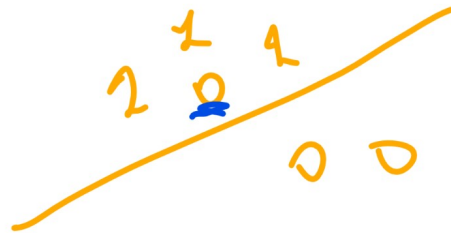
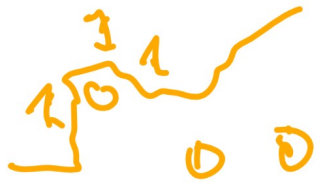
$$|3-1| + |6-1| = 7$$

3NN  
 ↓  
 1

tie breaking  
 def: use instance with smaller index.

# 3 Nearest Neighbor

## Quiz



nNN  
 ⇒ majority  
 vote

- Find the 3 Nearest Neighbor label for  $\begin{bmatrix} 3 \\ 3 \end{bmatrix}$  using Manhattan distance.

$x_1$	1	1	3	5	2
$x_2$	1	7	3	4	5
$y$	0	1	1	0	0

dist 4 6 0 3 3

- A: 0
- B: 1

C: I don't understand KNN.

# $K$ Fold Cross Validation

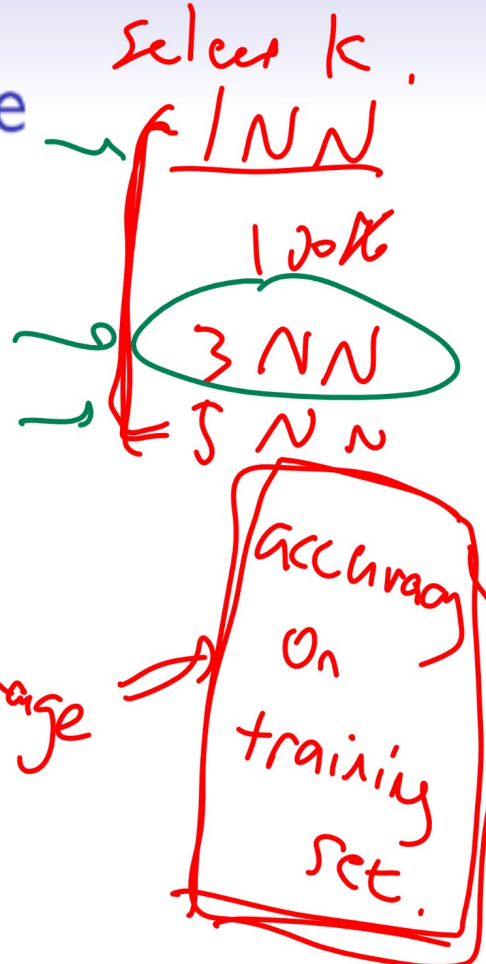
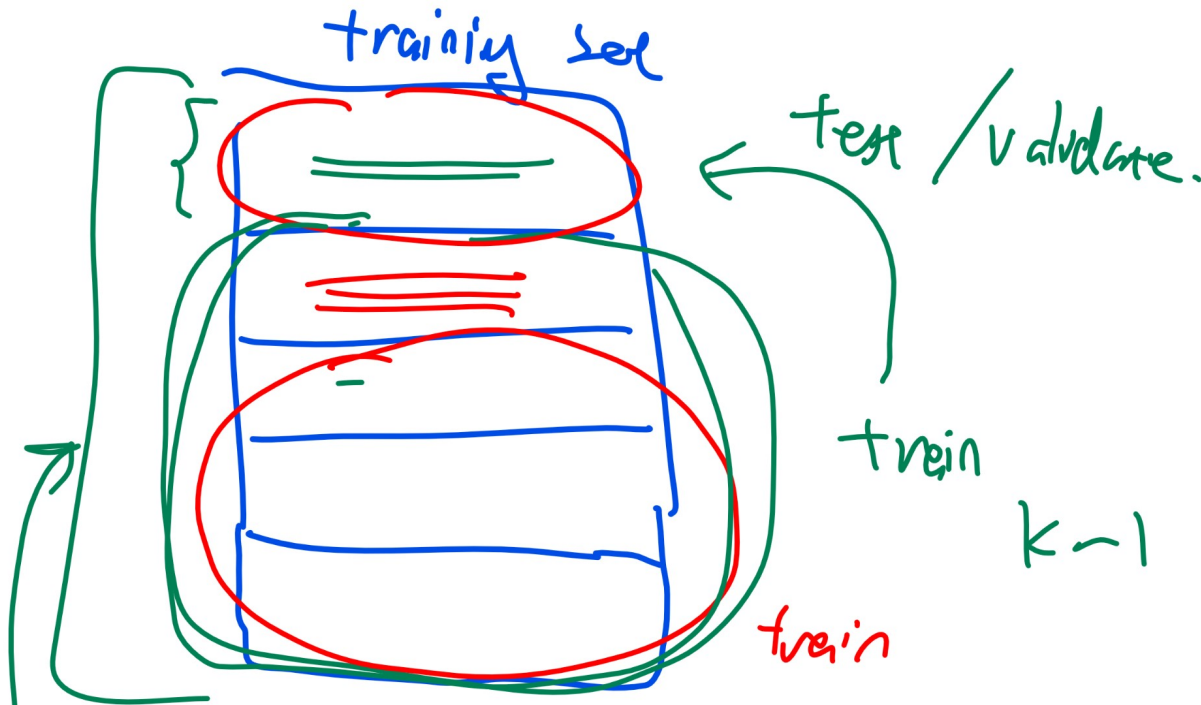
## Discussion

- Partition the training set into  $K$  groups.
- Pick one group as the validation set.
- Train the model on the remaining training set.
- Repeat the process for each of the  $K$  groups.
- Compare accuracy (or cost) for models with different hyperparameters and select the best one.



# 5 Fold Cross Validation Example

Discussion



1 fold < fold > fold



# Leave One Out Cross Validation

## Discussion

- If  $K = n$ , each time exactly one training instance is left out as the validation set. This special case is called Leave One Out Cross Validation (LOOCV).



# Cross Validation 2

## Quiz

Q3

- Given the following training data. What is the 10 fold cross-validation (LOOCV) accuracy if 1 nearest neighbor classifier with Manhattan distance is used?

x	1	1	2	2	3	3	4	4	5	5
y	1	2	3	3	2	2	3	3	2	1

- A : 20 percent, B: 40, C: 60, D: 80, E: 100

~~X~~ ~~\*~~ 3 3 2 2 3 3 *I don't understand LOOCV* ~~\*~~ ~~\*~~

$$\frac{6}{10} = 60\%$$