CS540 Introduction to Artificial Intelligence
Lecture 6

Young Wu
Based on lecture slides by Jerry Zhu and Yingyu Liang

June 6, 2019

Hat Game

Common knowledge

Quiz (Participation)

Contains information

- 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?

r1

- Everyone said no.
- Dad asked again: do you know the color of your hat?
- Everyone said no.

r2

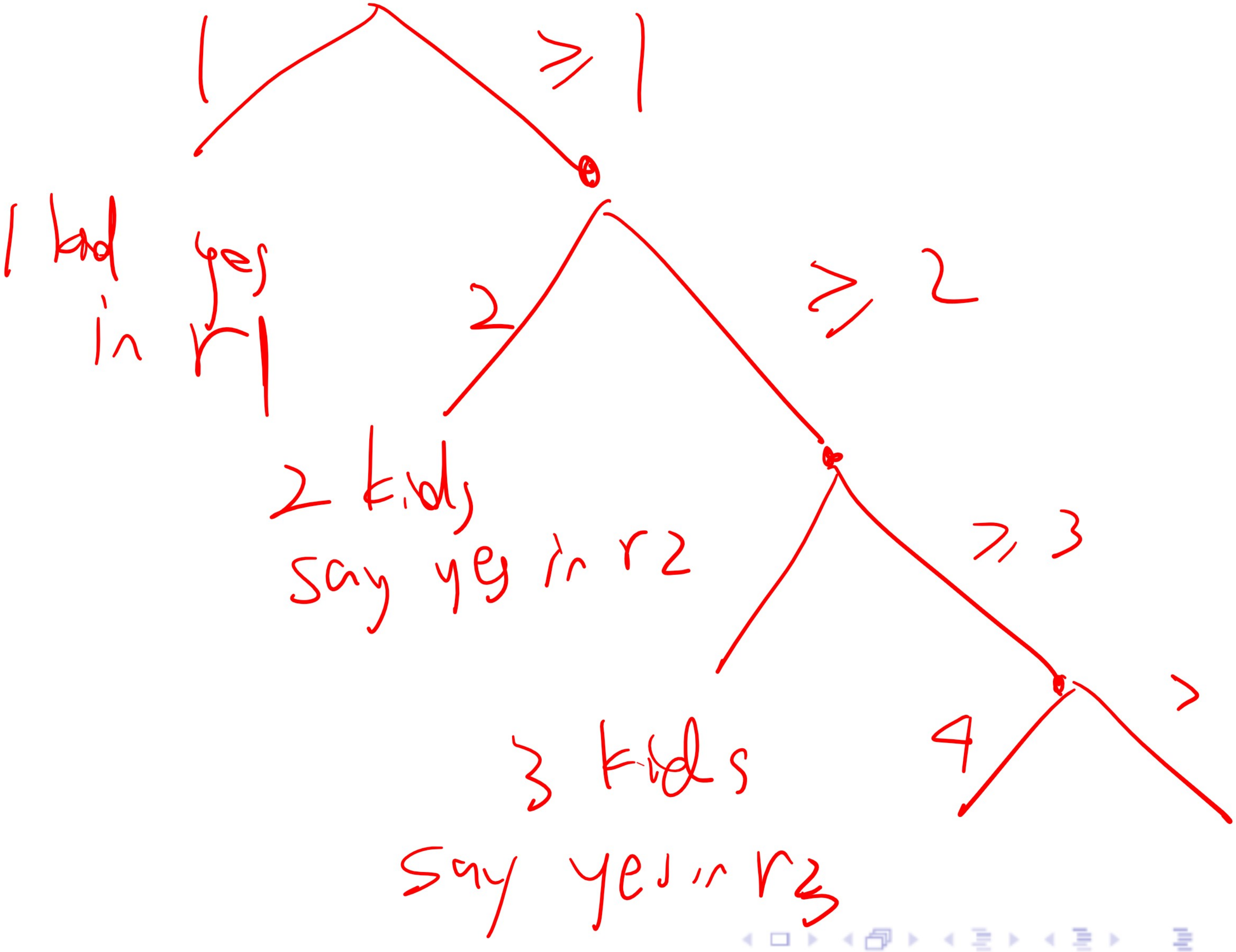
- Dad asked again: do you know the color of your hat?
- Some kids (at least one) said yes.

r3

- No one lied. How many kids are wearing green hats?
- A: 1... B: 2... [✓ C: 3...] D: 4... E: 5

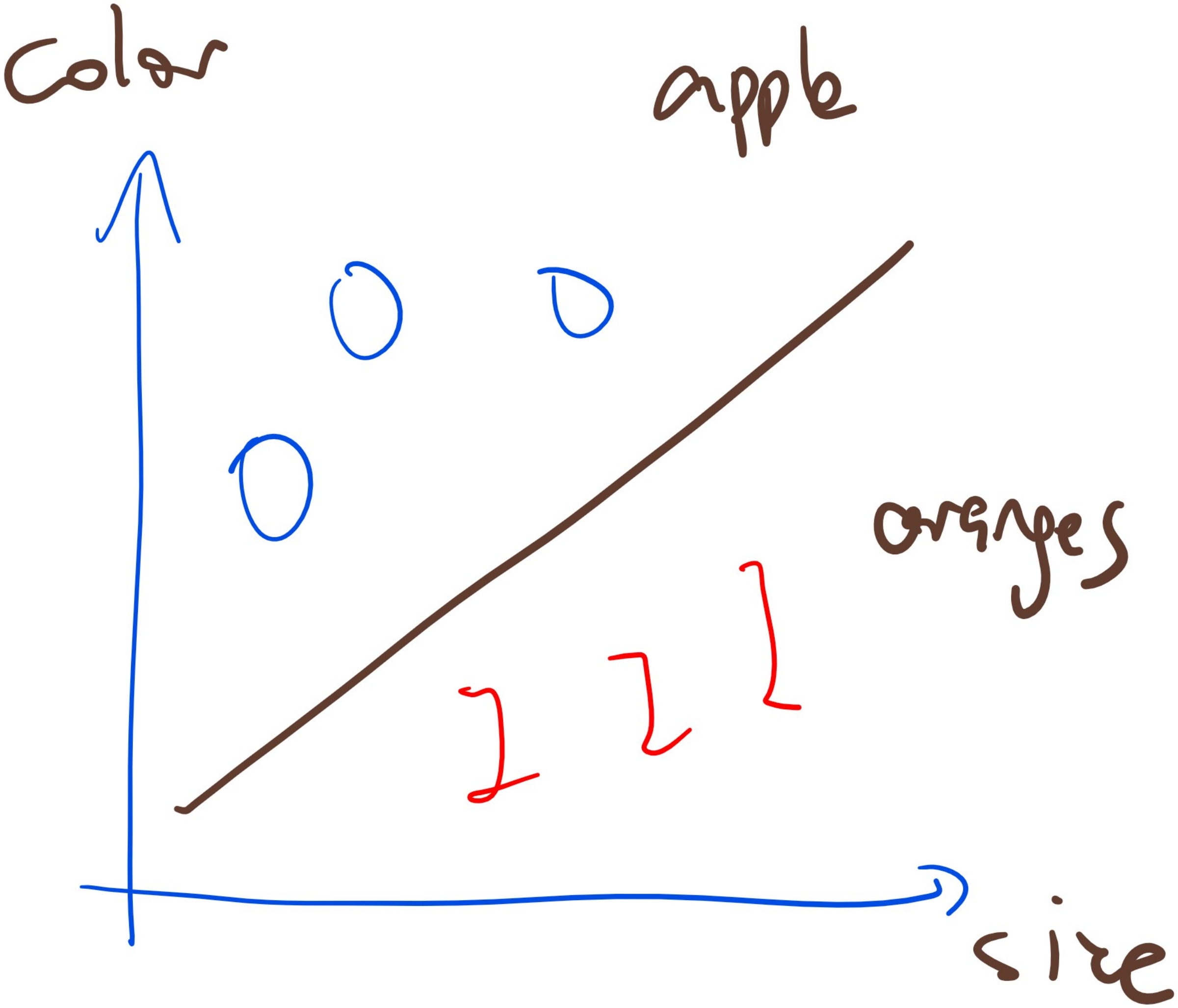
Hat Game Diagram

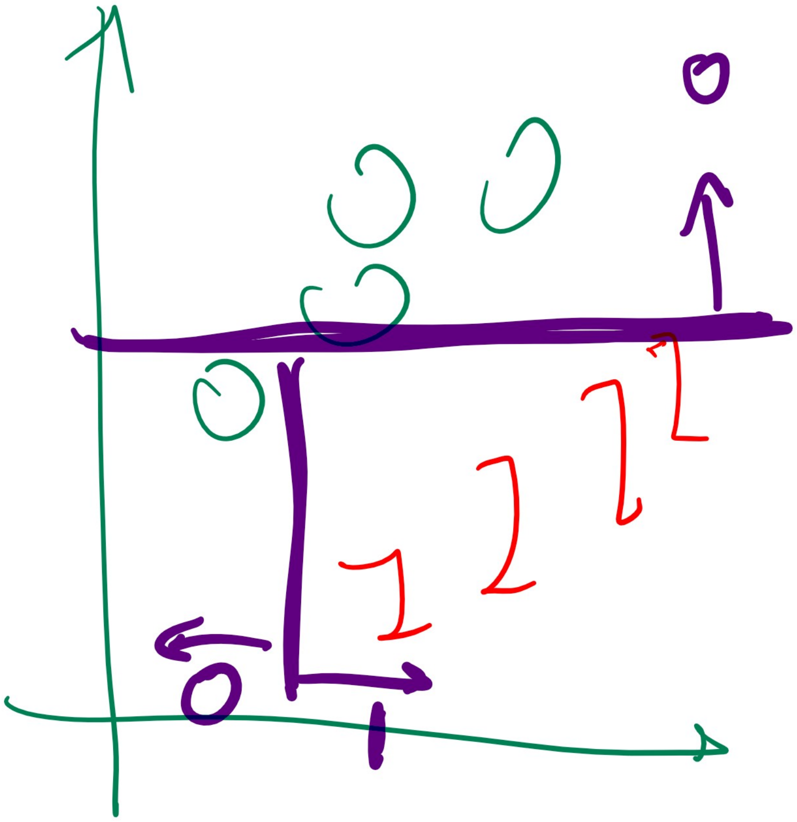
Discussion



Axes Aligned Decision Boundary

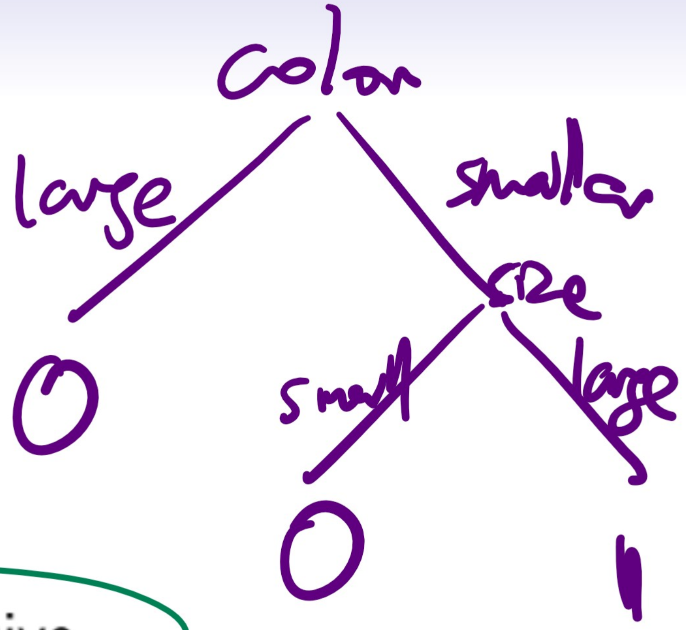
Motivation





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

opposite of information

- Entropy is the measure of uncertainty.
- 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:

P_0 $\frac{1}{P_0}$

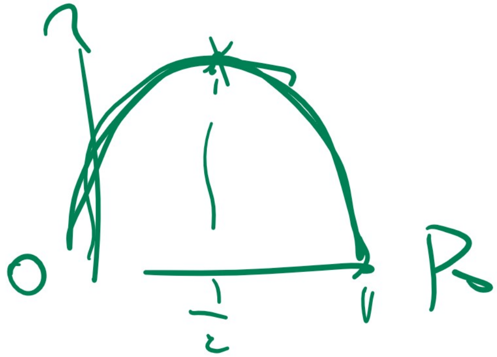
$$H(Y) = \underbrace{p_0 \underbrace{\log_2}\left(\frac{1}{p_0}\right) + p_1 \log_2\left(\frac{1}{p_1}\right)}$$

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

Measure of Uncertainty

Definition

- If $p_0 = 0$ and $p_1 = 1$, the entropy is 0, the outcome is certain, so there is no uncertainty.
- If $p_0 = 1$ and $p_1 = 0$, the entropy is 0, the outcome is also certain, so there is no uncertainty.
- If $p_0 = \frac{1}{2}$ and $p_1 = \frac{1}{2}$, the entropy is the maximum 1, the outcome is the most uncertain.



Entropy

Definition

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

$$H(Y) = \sum_{y=1}^{K} p_y \log_2 \left(\frac{1}{p_y} \right)$$

$$= - \sum_{y=1}^{K} p_y \log_2 (p_y)$$

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 is x, and $p_{y|x}$ as the fraction of the labels that are y among the ones with instance x.

$$H\left(Y|X=x\right) = -\sum_{y=1}^{K_Y} p_{y|x}\log_2\left(p_{y|x}\right)$$

$P_{0|5}$

$$H\left(Y|X\right) = \sum_{x=1}^{K_X} p_x H\left(Y|X=x\right)$$

fraction of instances with $x=5$, that has label 0

Aside: Cross Entropy

Definition

- Cross entropy measures the difference between two distributions.

$$H\left(Y,X\right) = -\sum_{z=1}^{K} \underline{p_{Y=z}} \log_2\left(\underline{p_{X=z}}\right)$$

- 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.

$$H\left(Y_i, A_i\right) = -y_i \log\left(a_i\right) - \left(1 - y_i\right)\log\left(1 - a_i\right)$$

Information Gain

Definition

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

$$\underline{I}(Y|X) = \underbrace{H(Y)}_{\text{uncertainty of } Y} - \underbrace{H(Y|X)}_{\text{uncertainty of } Y \text{ if } X \text{ value is given}}.$$

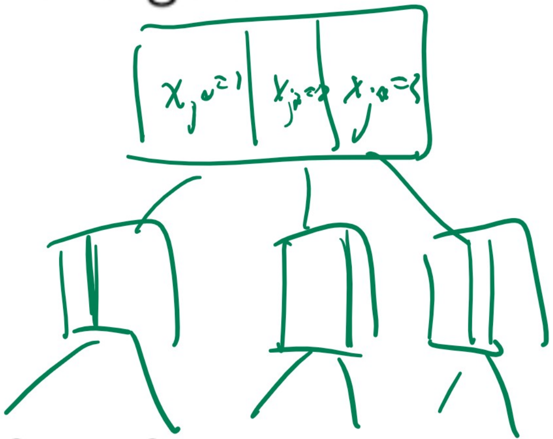
- The larger than information gain, the larger the reduction in uncertainty, and the better predictor the feature is.

Splitting Discrete Variables

Definition

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

$$\arg\max_j I\left(Y|X_j\right)$$



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

$$\left\{(x_i, y_i) : x_{ij} = 1\right\}, \left\{(x_i, y_i) : x_{ij} = 2\right\}, \ldots, \left\{(x_i, y_i) : x_{ij} = K_{X_j}\right\}$$

Splitting Continuous Variables

Definition

- Continuous variables can be uniformly split into K_X categories.
- In practice, all possible binary splits of the continuous variables are constructed, and the one that yields the highest information gain is used.

$$\mathbb{1}_{\{x_j > x_{1j}\}}, \mathbb{1}_{\{x_j > x_{2j}\}}, \ldots, \mathbb{1}_{\{x_j > x_{nj}\}}$$

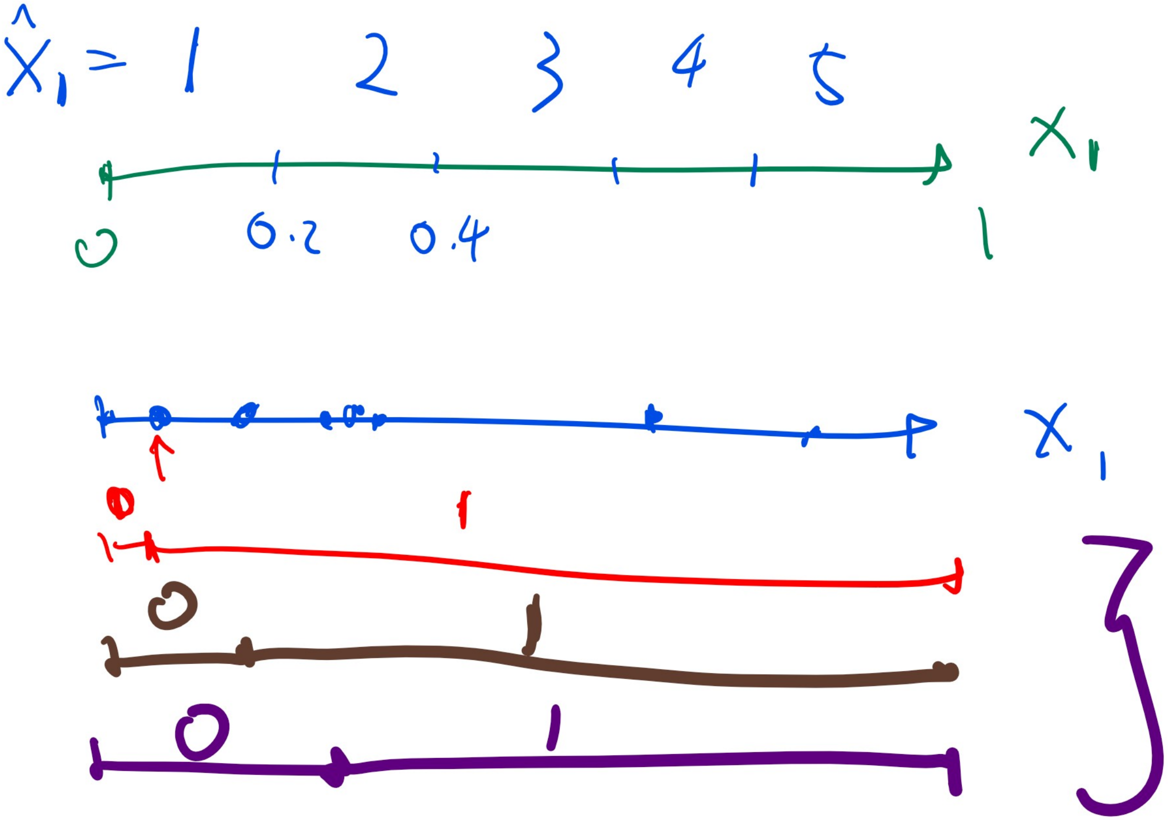
- One of the above binary features is used in place of the original continuous variable x_j.

Splitting Continuous Variables Diagram

Definition

HW3

in practice



try all

find the one with largest info gain.

ID3 Algorithm (Iterative Dichotomiser 3), Part I

Algorithm

- Input: instances: $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, feature j is split into K_j categories and y has K categories
- Output: a decision tree
- Start with the complete set of instances $\{x_i\}_{i=1}^n$.
- Suppose the current subset of instances is $\{x_i\}_{i \in S}$, find the information gain from each feature.

$$I(Y|X_j) = H(Y) - H(Y|X_j)$$

ID3 Algorithm (Iterative Dichotomiser 3), Part II

Algorithm

$$H(Y) = -\sum_{y=1}^{K} \frac{\#(Y=y)}{\#(Y)} \log\left(\frac{\#(Y=y)}{\#(Y)}\right)$$

$$H(Y|X_j) = -\sum_{x=1}^{K_j}\sum_{y=1}^{K} \frac{\#(Y=y, X_j=x)}{\#(Y)} \log\left(\frac{\#(Y=y, X_j=x)}{\#(X_j=x)}\right)$$

- Find the more informative feature $j^\star$.

$$j^\star = \arg\max_j I(Y|X_j)$$

ID3 Algorithm (Iterative Dichotomiser 3), Part III

Algorithm

- Split the subset S into $K_{j^\star}$ subsets.

$$\begin{aligned} S_1 &= \{(x_i, y_i) \in S : x_{ij^\star} = 1\} \\ S_2 &= \{(x_i, y_i) \in S : x_{ij^\star} = 2\} \\ &\dots \\ S_{K_{X_{j^\star}}} &= \left\{(x_i, y_i) \in S : x_{ij^\star} = K_{X_{j^\star}}\right\} \end{aligned}$$

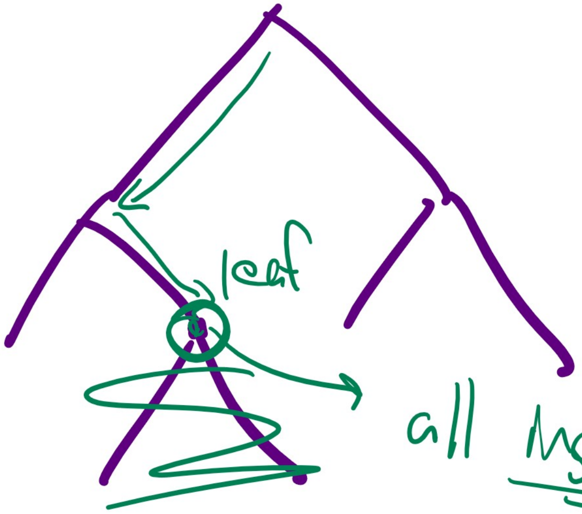
- Recurse over the subsets until $\underline{p_y = 1}$ for some y on the subset.

Stop

Pruning

Discussion

avoid overfitting



all instances here have the same label

validation set

- Use the validation set to prune subtrees by making them a leaf. The leaf has label equal to the majority of the train examples reaching this subtree.

$\log a^b = b \log a$ 12

Entropy

Quiz (Graded)

$\underbrace{\log \frac{1}{p}} = \log p^{-1} = \underbrace{-\log p}$

$\log_2 8 = 3$

$\log_2 \frac{1}{4} = -\log_2 4$

$\log_2 4 = 2^? = 4 \Rightarrow ? = 2$

- Fall 2010 Final Q10
- 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?

1 2 3

$\log \frac{1}{p} = -\log(p)$ $Y = 1, 2, 3$

- A: 0.75
- B: 1
- [✓ C: 1.5]
- D: 1.75
- E: None of the above.

$$H(Y) = \underbrace{-p_1 \log_2 p_1 - p_2 \log_2 p_2 - p_3 \log_2 p_3}$$

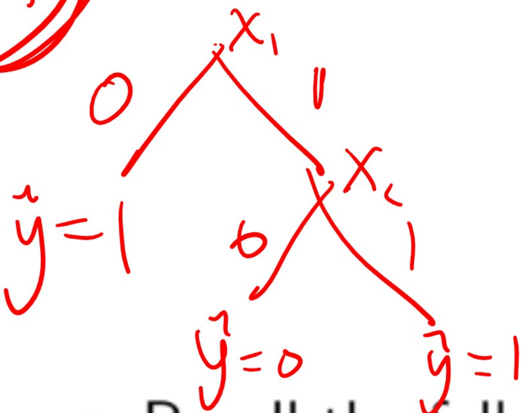
$$+\frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 \quad \bigcirc$$

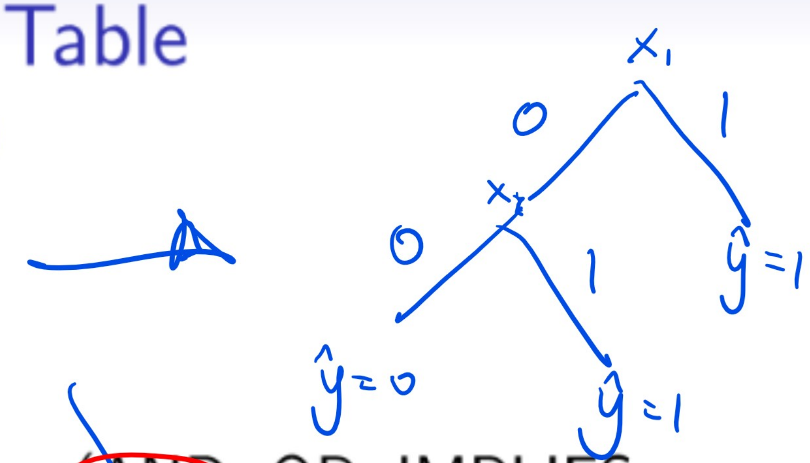
$$0.5 + 0.5 + 0.5$$

14

Decision Tree, Table

Quiz (Graded)





- Recall the following logical operators (AND, OR, IMPLIES, IF).

A. B. C. D.

x_1	x_2	$x_1 \wedge x_2$	$x_1 \vee x_2$	$x_1 \Rightarrow x_2$	$x_1 \Leftarrow x_2$
1	1	1	1	1	1
1	0	0	1	0	1
0	1	0	1	1	0
0	0	0	0	1	1

Decision Tree

Quiz (Graded)

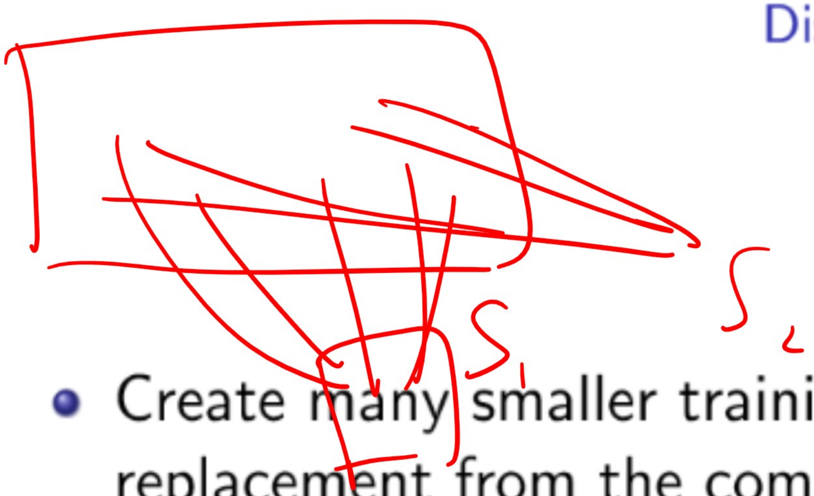
- Fall 2009 Midterm Q2
- Which expression is represented by the decision tree:

$$x_1 \begin{cases} T & \hat{y} = T \\ F & x_2 \begin{cases} T & \hat{y} = T \\ F & \hat{y} = F \end{cases} \end{cases}$$

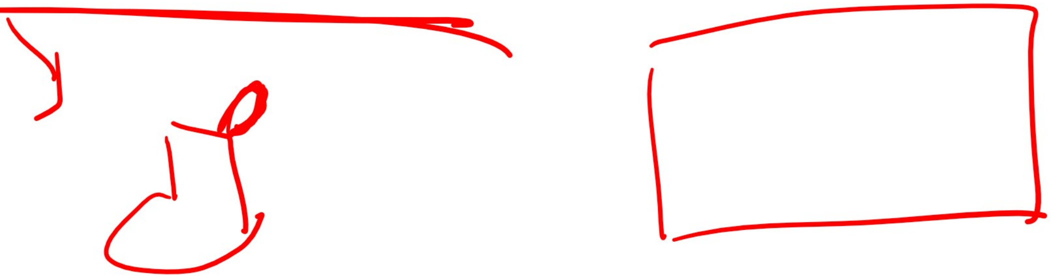
- A: $x_1 \wedge x_2$ (AND)
- B: $x_1 \vee x_2$ (OR)
- C: $x_1 \Rightarrow x_2$ (IMPLIES)
- D: $x_1 \Leftarrow x_2$ (IF)
- E: None of the above.

Bagging

Discussion



- Create many smaller training sets by sampling with replacement from the complete training set.
- Train different decision trees using the smaller training sets.
- Predict the label of new instances by majority vote from the decision trees.
- This is called bootstrap aggregating (bagging).



Random Forrest

Discussion

- When training the decision trees on the smaller training sets, only a random subset of the features are used. The decision trees are created without pruning.
- This algorithm is called random forests.

Boosting

Discussion

- The idea of boosting is to combine many weak decision trees, for example, decision stumps, into a strong one.
- Decision trees are trained sequentially. The instances that are classified incorrectly by previous trees are made more important for the next tree.

Adaptive Boosting, Part I

Discussion

- The weights w for the instances are initialized uniformly.

$$w = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$$

- In each iteration, a decision tree f_k is trained on the training instances weighted by w.
- The weights are updated according to the error made by f_k.

$$w_i = w_i \frac{\varepsilon}{1-\varepsilon} \mathbb{1}_{\{f_k(x_i)=y_i\}}$$

$$\varepsilon = \sum_{i=1}^{n} w_i \mathbb{1}_{\{f_k(x_i)\neq y_i\}}$$

Adaptive Boosting, Part II

Discussion

- The weights are then normalized (to have sum $= 1$) and the weights for the trees z_j are updated.

$$z_j = \log \frac{1 - \varepsilon}{\varepsilon}$$

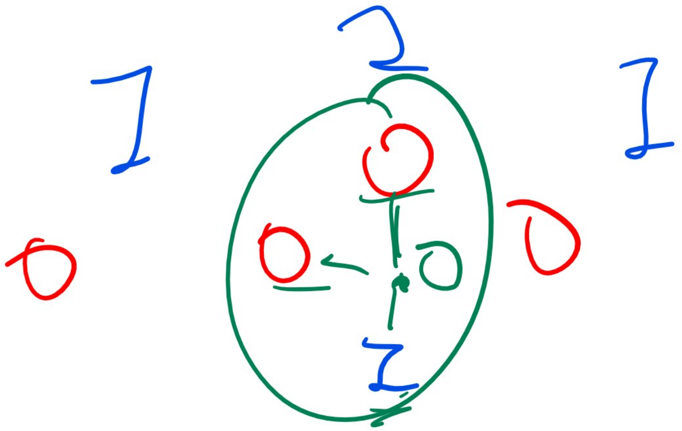
- The label of a new test instance x_i is the z weighted majority of the labels produced by all K trees: $f_1(x_i), f_2(x_i), ..., f_K(x_i)$.

8:05

K Nearest Neighbor

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.



3NN

$\sqrt{x_1^2 + x_2^2}$

Distance Function

Definition

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

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

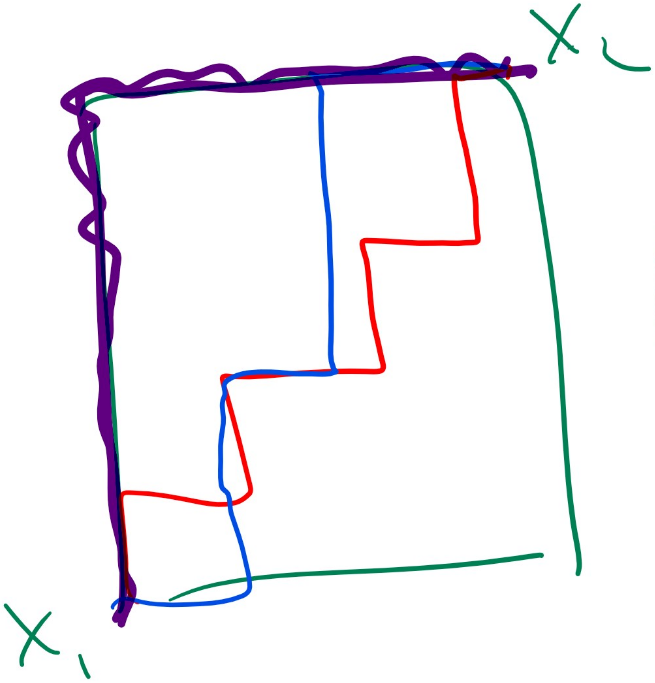
- An example is Manhattan distance.

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

L1 - norm

Manhattan Distance Diagram

Definition



$$|X_{11} - X_{21}|$$

$$+|X_{21} - X_{22}|$$

M distance

P Norms

Definition

- Another group of examples is the p norms.

$$\rho\left(x, x'\right) = \left(\sum_{j=1}^{m} \left|x_j - x_j'\right|^p\right)^{\frac{1}{p}}$$

- $p = 1$ is the Manhattan distance.
- $p = 2$ is the Euclidean distance.
- $p = \infty$ is the sup distance, $\rho\left(x, x'\right) = \underbrace{\max_{i=1,2,\ldots,m} \left\{\left|x_j - x_j'\right|\right\}}$.
- p cannot be less than 1.

K Nearest Neighbor

Algorithm

- Input: instances: $\{x_i\}_{i=1}^{n}$ and $\{y_i\}_{i=1}^{n}$, and a new instance $\hat{x}$.
- Output: new label $\hat{y}$.
- Order the training instances according to the distance to $\hat{x}$.

$$\rho\left(\hat{x}, x_{(i)}\right) \leqslant \rho\left(\hat{x}, x_{(i+1)}\right), i = 1, 2, ..., n-1$$

- Assign the majority label of the closest k instances.

$$\hat{y} = \text{mode}\left\{y_{(1)}, y_{(2)}, ..., y_{(k)}\right\}$$

1 Nearest Neighbor

Quiz (Graded)

16

- Spring 2018 Midterm Q7
- Find the 1 Nearest Neighbor label for $\overset{x^*}{\begin{bmatrix}3\\6\end{bmatrix}}$ using Manhattan distance.

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

NN

$$|x^*_1 - x_1| + |x^*_2 - x_2|$$

3NN ⇒ 1

- [✓ A: 0]
- B: 1
- C, D, E: Don't choose.

5 Nearest Neighbor

Quiz (Graded)

- Spring 2018 Midterm Q7
- Find the 5 Nearest Neighbor label for $\begin{bmatrix}3\\6\end{bmatrix}$ using Manhattan distance. $K = n$

majority classifier

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

K-D tree

- [✓ A: 0]
- B: 1
- C, D, E: Don't choose.