CS540 Introduction to Artificial Intelligence
Lecture 6

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

June 5, 2019

Hat Game

Quiz (Participation)

Common knowledge

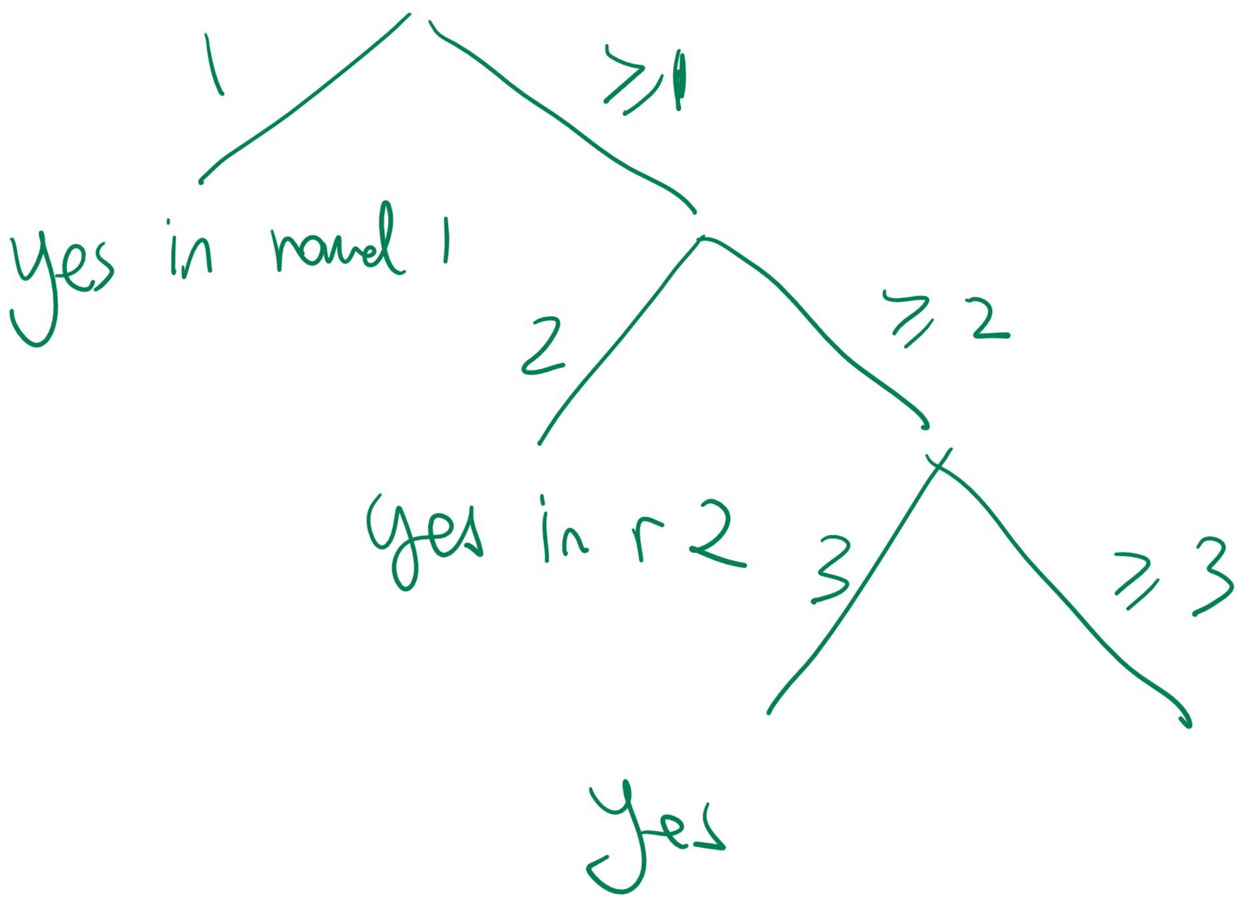
- 5 kids are wearing either green or red hats in a party: they can see every other kid's hat but not their own.
- Dad said to everyone: at least one of you is wearing green hat.
- Dad asked everyone: do you know the color of your hat?
- Everyone said no.
- Dad asked again: do you know the color of your hat?
- Everyone said no.
- 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

information ?

everyone knew this at the beginning

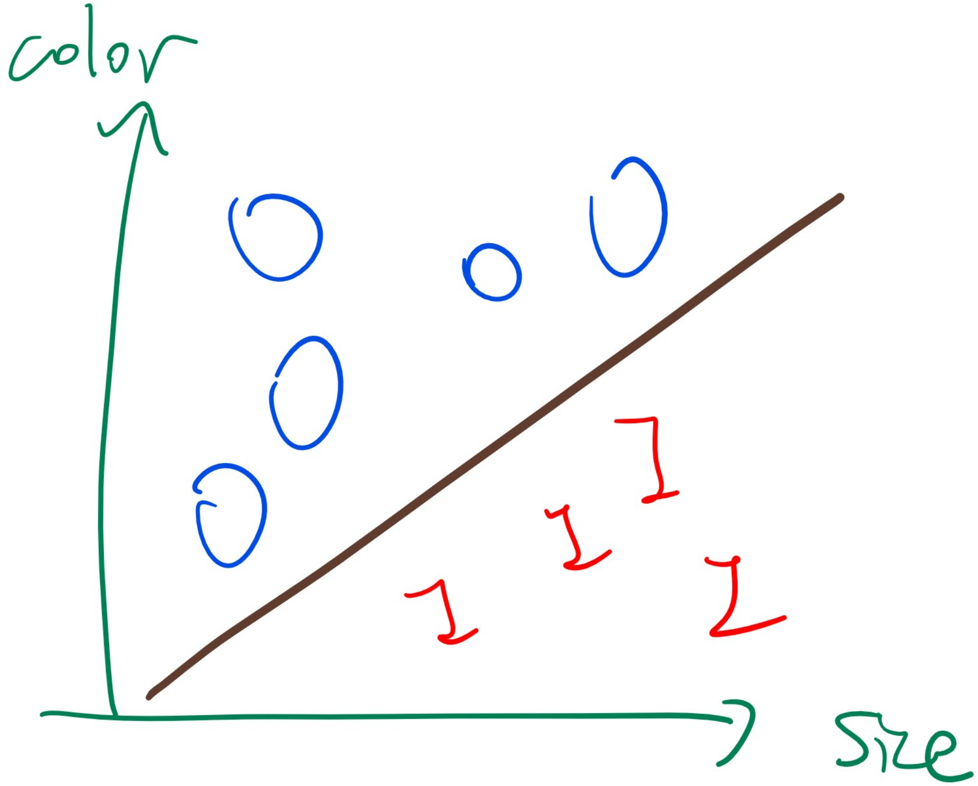
Hat Game Diagram

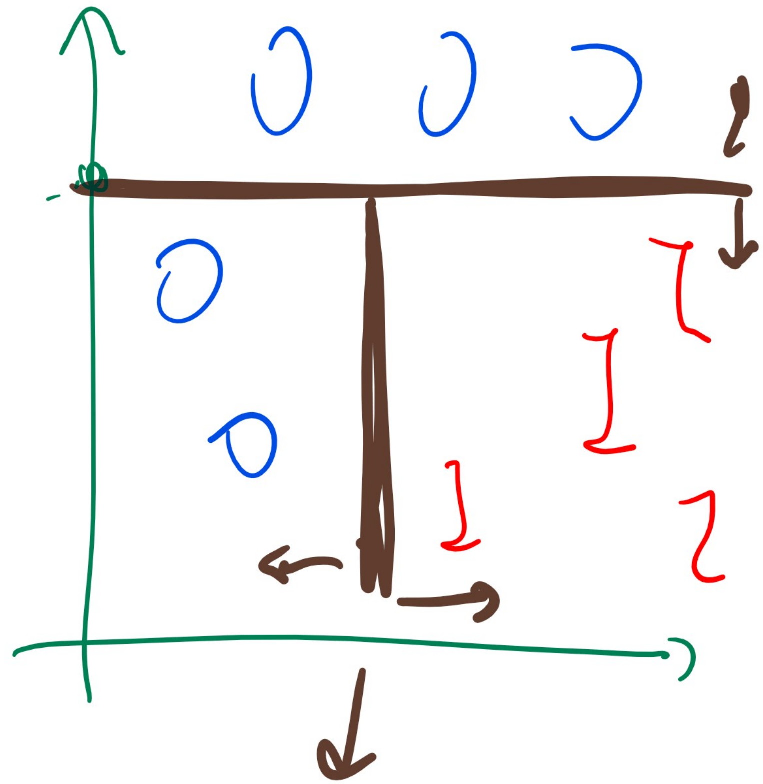
Discussion



Axes Aligned Decision Boundary

Motivation





decision trees

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:

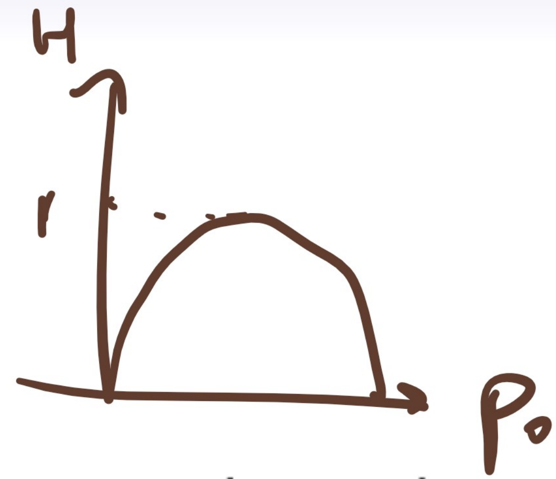
prob ↑ info ↓

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

convert into bits.

$$E\left[\log_2\left(\frac{1}{p}\right)\right]$$

Measure of Uncertainty



Definition

assume $0 \lg 0 = 0$

- 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 $y \in \{1, 2, \ldots, K\}$, 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.

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

entropy of conditional prob

fraction y given x.

$$H(Y|X) = \underbrace{\sum_{x=1}^{K_X} p_x H(Y|X=x)}_{p_x \text{ weighted average}}$$

Aside: Cross Entropy

Definition

- Cross entropy measures the difference between two distributions.

log

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

- 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.

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

Information Gain

Definition

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

$$I(Y|X) = \underbrace{H(Y) - H(Y|X)}_{\text{reduction in uncertainty}}.$$

X is provided

reduction in uncertainty if X is provided

- 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.

$$j^* = \arg\max_j I\left(Y|X_j\right)$$

feature with max info gain.

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

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

$x_{ij^*} = 1$	$x_{ij^*} = 2$	$x_{ij^*} = 3$

Splitting Continuous Variables

Definition

HW3

set arbitrarily

- 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.

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

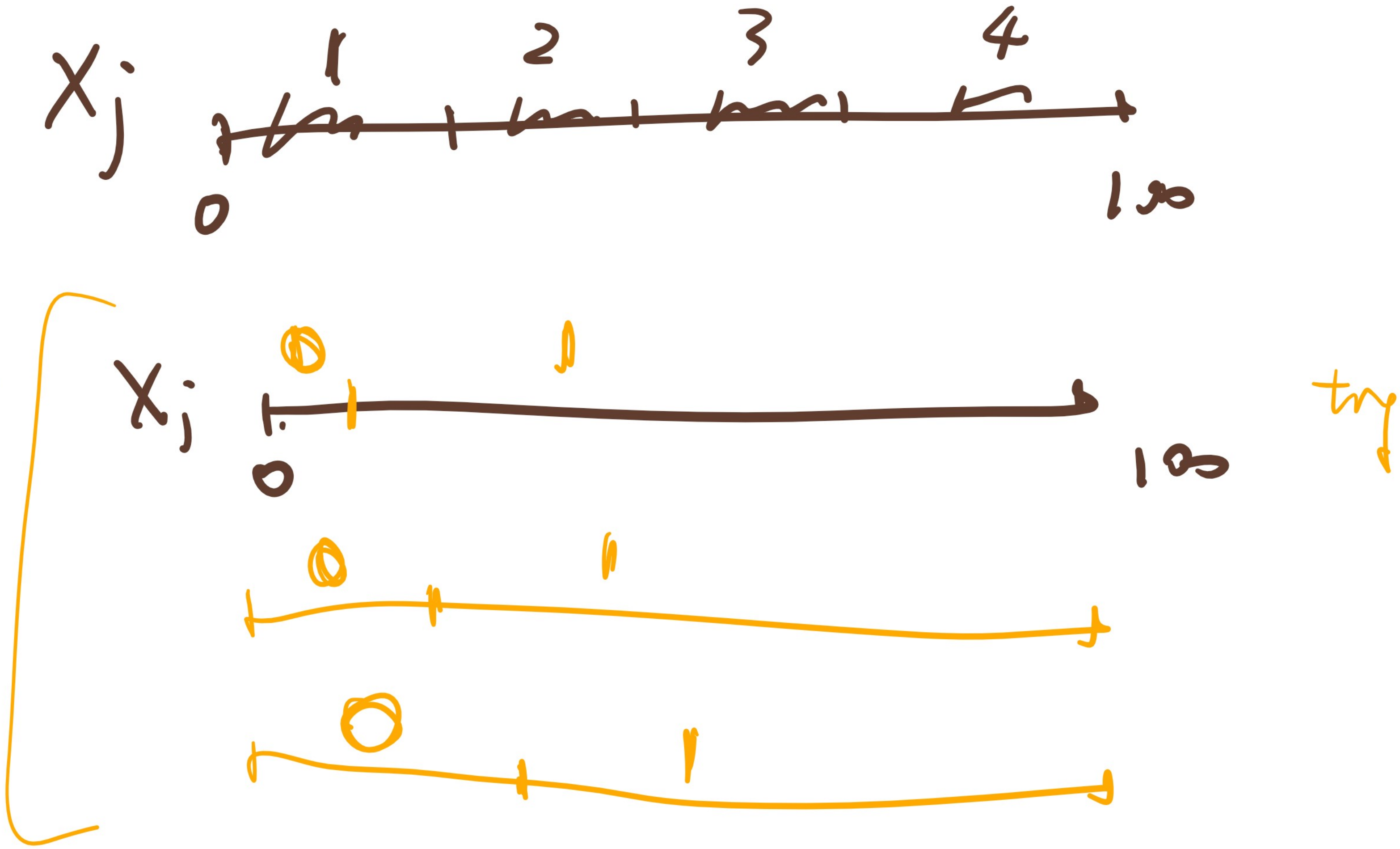
- One of the above binary features is used in place of the original continuous variable x_j.

Splitting Continuous Variables Diagram

Definition

HW3.

in practice



find max info gain ; repeat for all possible split

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

count

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

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

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

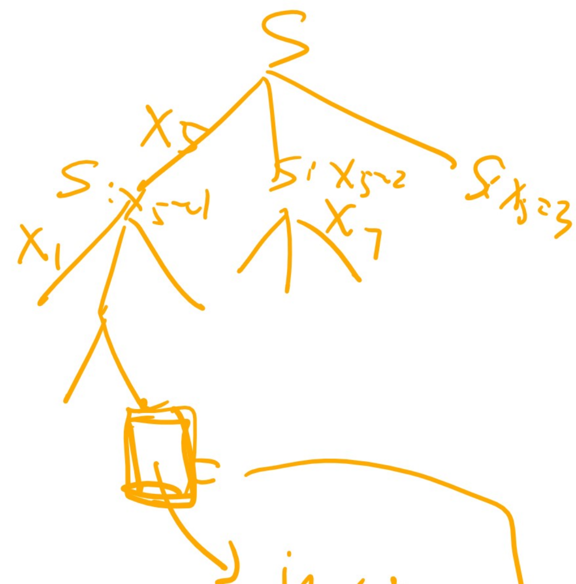
$$j^\star = \arg\max_j I(Y|X_j)$$

ID3 Algorithm (Iterative Dichotomiser 3), Part III

Algorithm

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

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



in same class

same label

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

Pruning

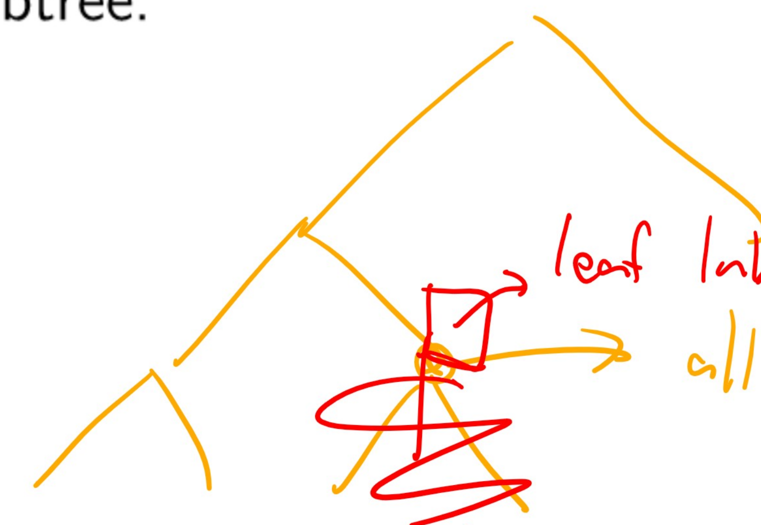
Discussion

avoid overfitting

- 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.

Simple tree

$\Longrightarrow$ no overfit



label of subset

leaf label = majority

all labels in validation is 1

cut branches

Entropy

Quiz (Graded)

Q3

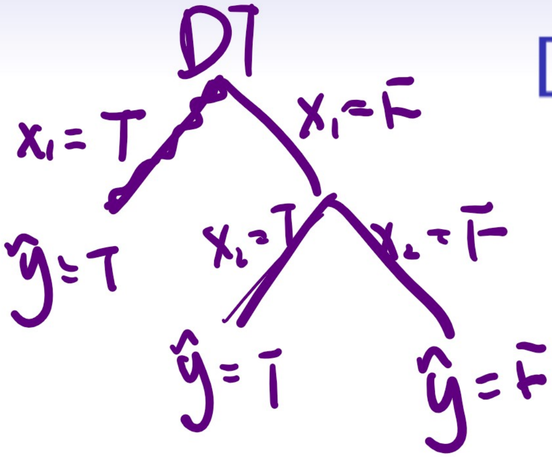
- Fall 2010 Final Q10
- Running from You-Know-Who, Harry enters the CS building on the 1st floor. He flips a fair coin: if it is heads he hides in room 1325; otherwise, he climbs to the 2nd floor. In that case he flips the coin again: if it is heads he hides in CSL; otherwise, he climbs to the 3rd floor and hides in 3331. What is the entropy of Harry's location?
- A: 0.75
- B: 1
- C: 1.5
- D: 1.75
- E: None of the above.

50%, 50%

$Y = 1, 2, 3$

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

$\overset{0.5}{\cancel{-\frac{1}{2}\log(\frac{1}{2})}} \overset{0.5}{\cancel{-\frac{1}{4}\log(\frac{1}{4})}} \overset{0.5}{\cancel{-\frac{1}{4}\log\frac{1}{4}}}$



Decision Tree, Table

Quiz (Graded)

T = 1
F = 0

- 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	$\boxed{1}$	1	1
1	0	0	$\boxed{1}$	$\cancel{0}$	1
0	1	0	$\boxed{1}$	1	0
0	0	0	$\boxed{0}$	1	1

Decision Tree

Quiz (Graded)

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

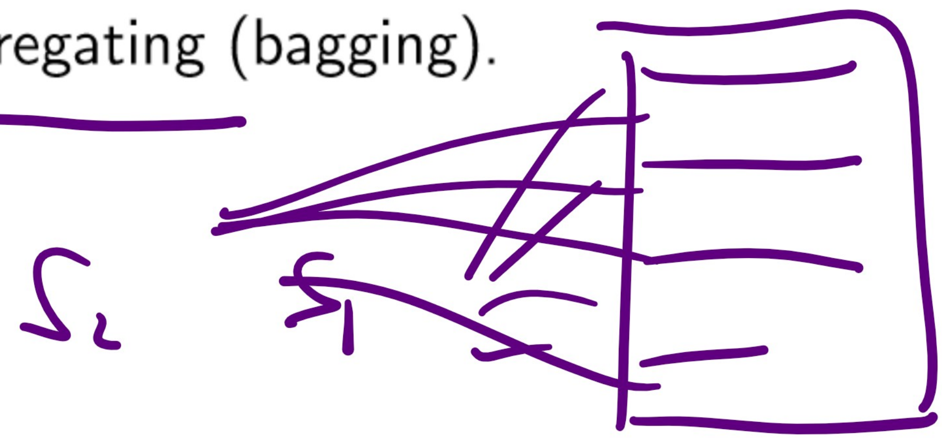
$$x_1 \begin{cases} T & \hat{y} = T \\ F & x_2 \begin{cases} T & \hat{y} = T \\ F & \hat{y} = F \end{cases} \end{cases}$$

- 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

random subsample of instances
and feature

- 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

Adaboost.

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.

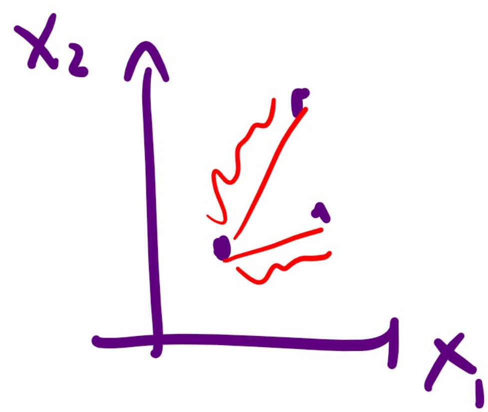
$$z_j = \log \frac{1 - \varepsilon}{\varepsilon}$$

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

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.



Distance Function

Definition

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

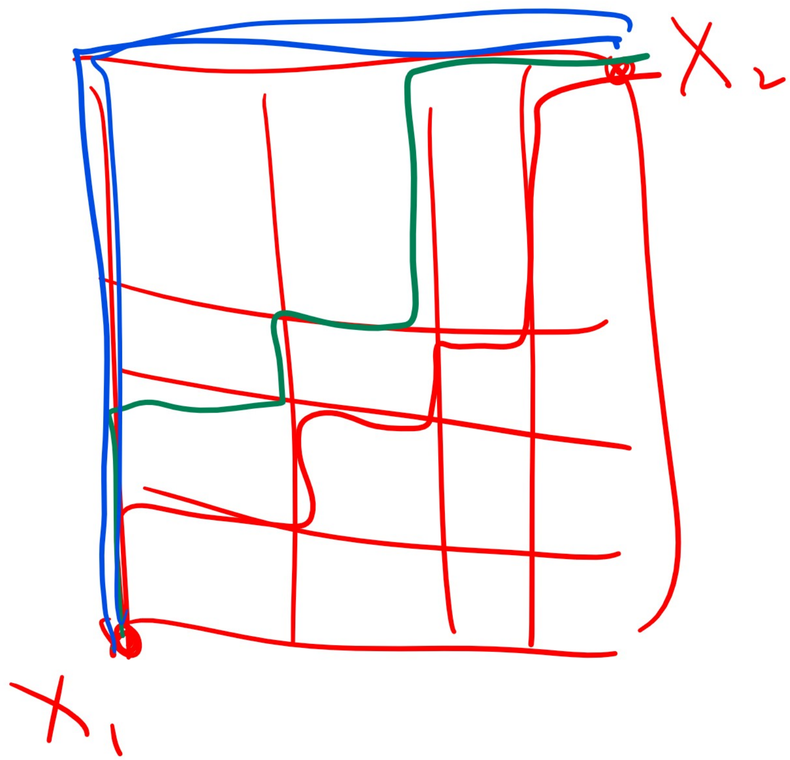
$$\rho\left(x, x'\right) = \left\|x - x'\right\|_2 = \sqrt{\sum_{j=1}^{m}\left(x_j - x'_j\right)^2}$$

- An example is Manhattan distance.

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

Manhattan Distance Diagram

Definition



Manhattan distance

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

$+ |X_{12} - X_{22}|$

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. ← often used.
- $p = 2$ is the Euclidean distance.
- $p = \infty$ is the sup distance, $\rho\left(x, x'\right) = \max_{i=1,2,\ldots,m} \left\{\left|x_j - x'_j\right|\right\}$.
- 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)

7

$|x_1^* - x_1| + |x_2^* - x_2|$

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

$\begin{pmatrix} x_1^* \\ x_2^* \end{pmatrix}$

x_1	1	1	3	5	2
x_2	1	7	3	4	5
y	0	1	1	0	0
	>4	3	3	4	$\boxed{2}$

- $\boxed{\text{A: 0}}$
- B: 1
- C, D, E: Don't choose.

5 Nearest Neighbor - Quiz (Graded)

8

- Spring 2018 Midterm Q7
- Find the 5 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
- C, D, E: Don't choose.