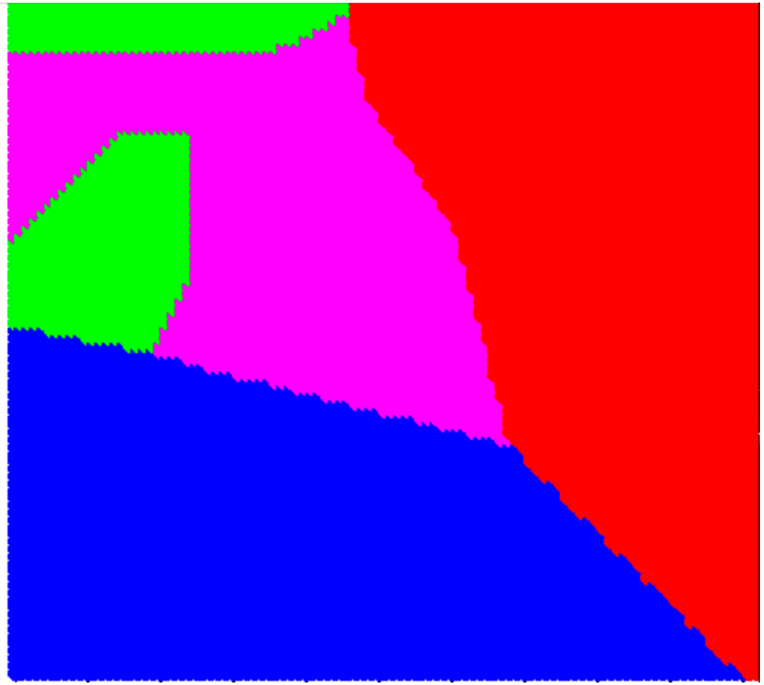
Decision Trees

Announcements

HW6 and P6 will be released soon

Recap on the K-NN algorithm

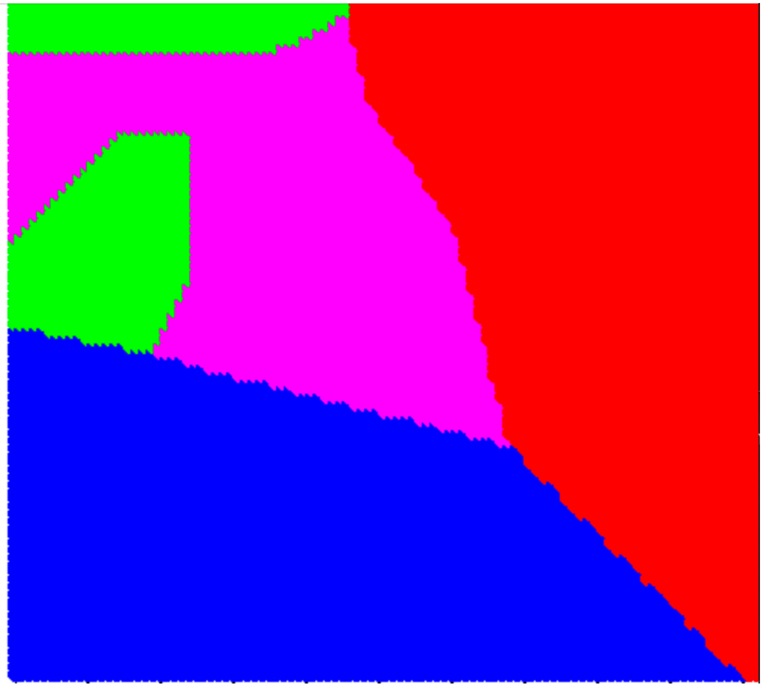
K-NN can have complicated nonlinear decision boundaries



[1-NN decision boundary in prelim]

Recap on the K-NN algorithm

K-NN can have complicated nonlinear decision boundaries



[1-NN decision boundary in prelim]

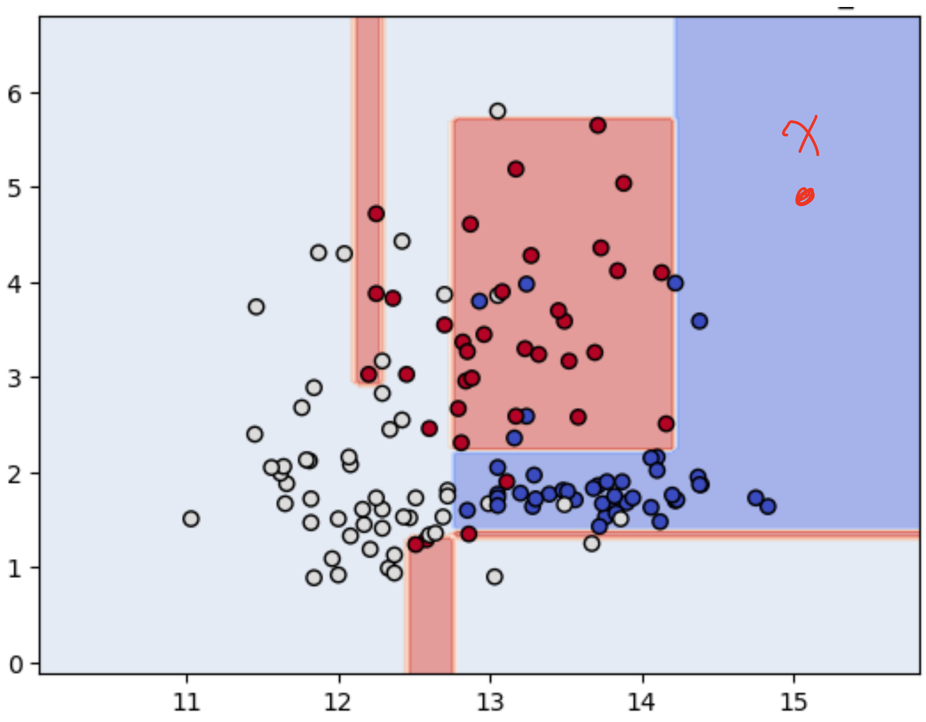
k-NN is expensive in computation and memory

Objective today

Decision tree — more efficient algorithm that
(1) splits space into regions with the same label, (2) is very fast in test time

Objective today

Decision tree — more efficient algorithm that
(1) splits space into regions with the same label, (2) is very fast in test time



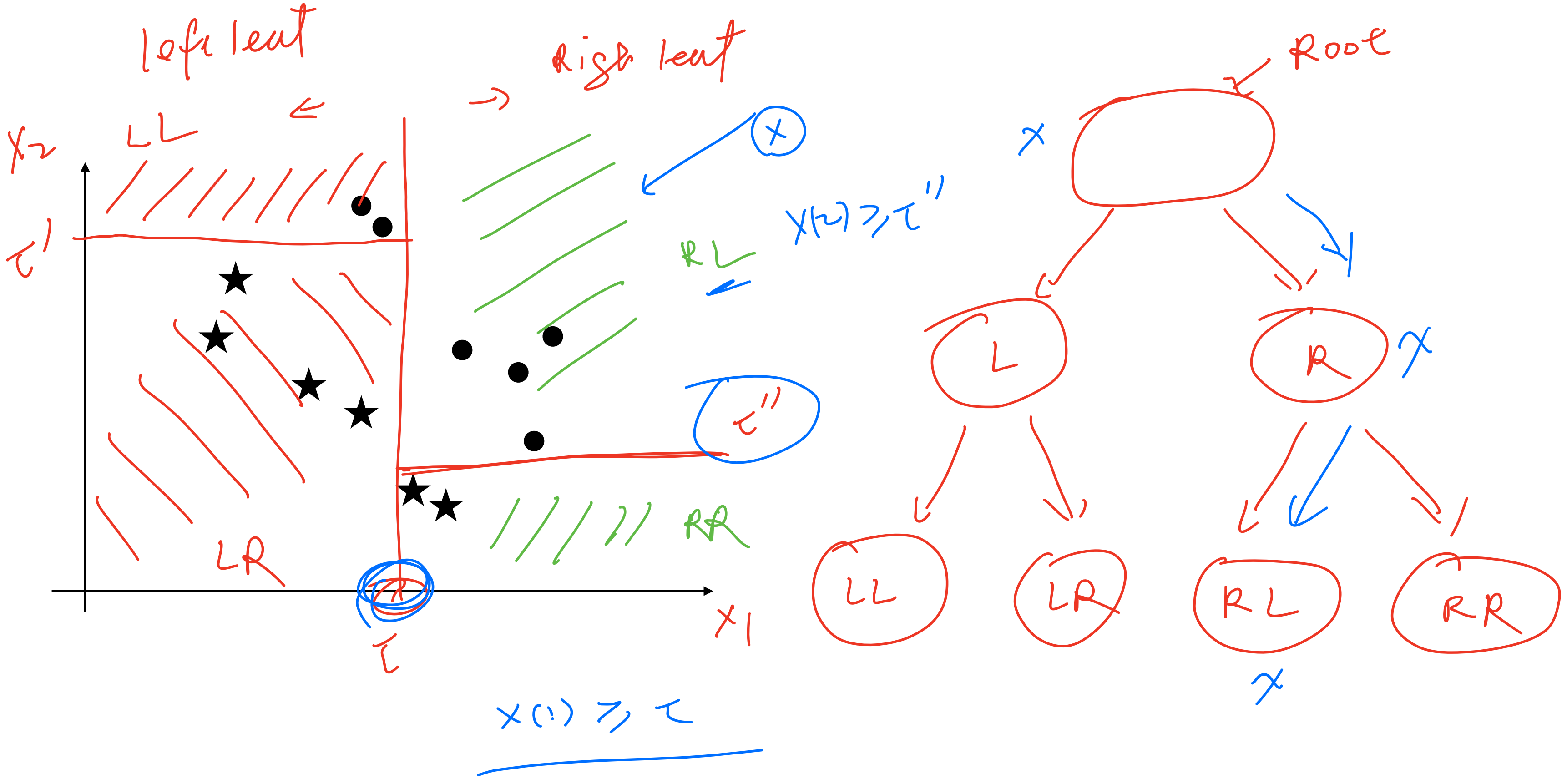
Outline of Today

1. Decision tree in classification

2. Decision tree in regression

3. Demos of decision tree

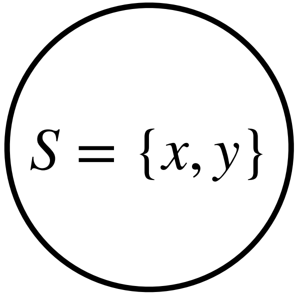
Overview of the Decision Tree algorithm

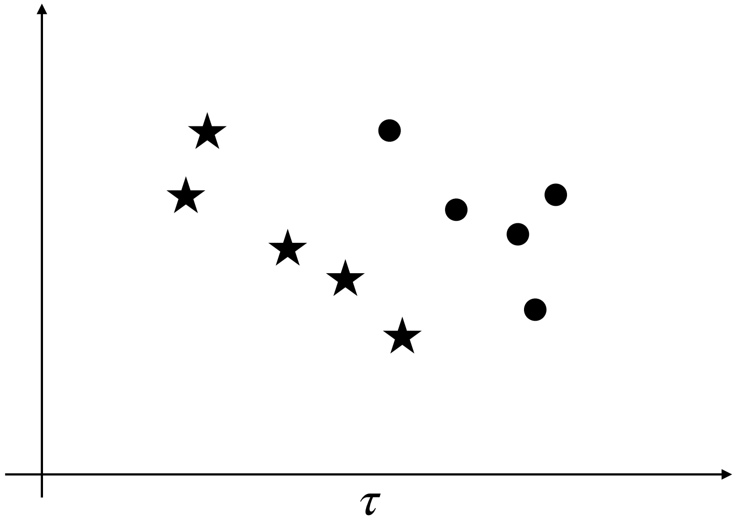


How to split a tree node

$X \in R$

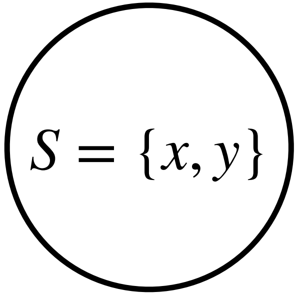
Consider k-class classification, i.e., $y \in \{1,2,\ldots,k\}$

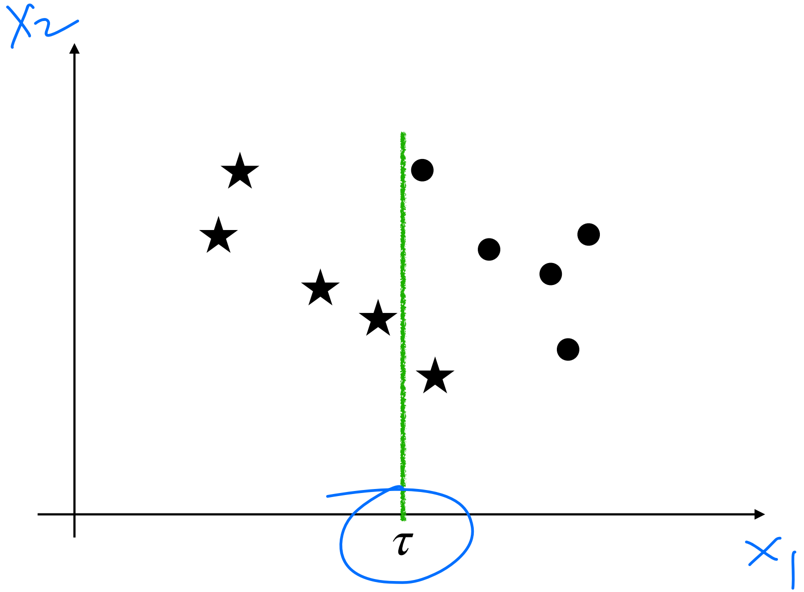




How to split a tree node

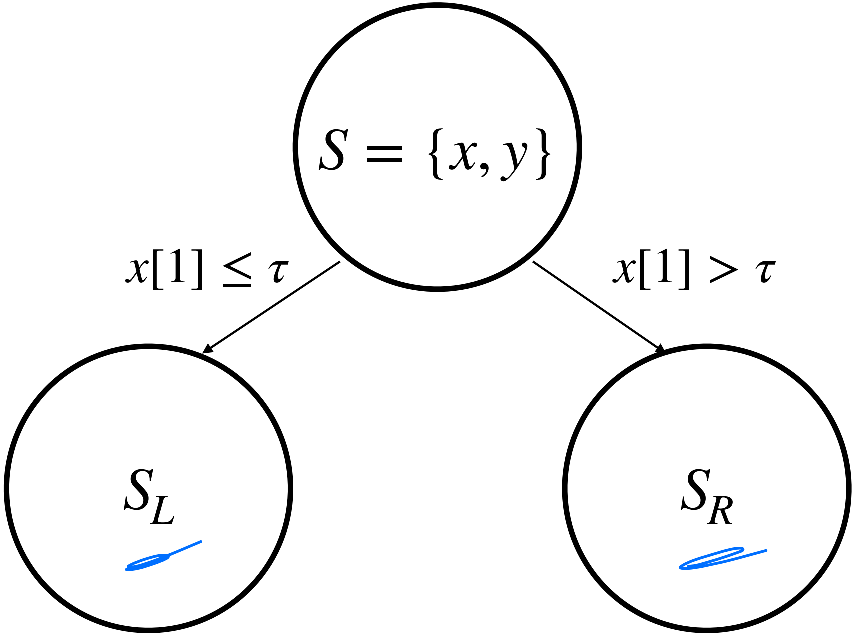
Consider k-class classification, i.e., $y \in \{1,2,\ldots,k\}$

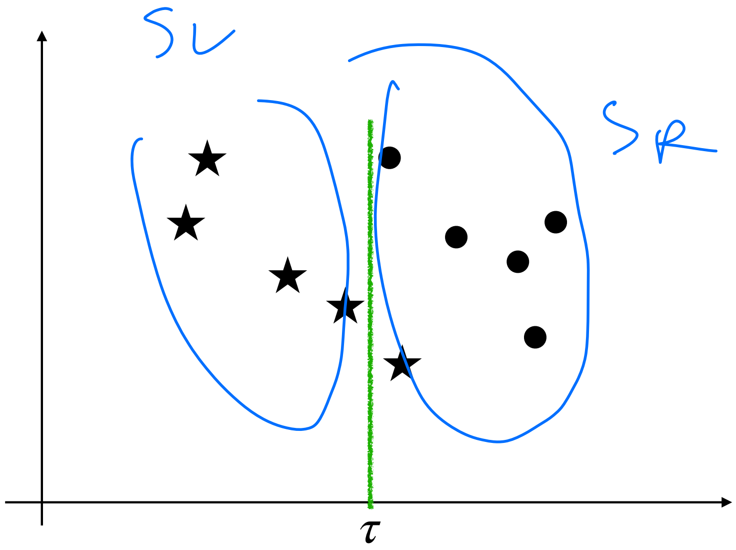




How to split a tree node

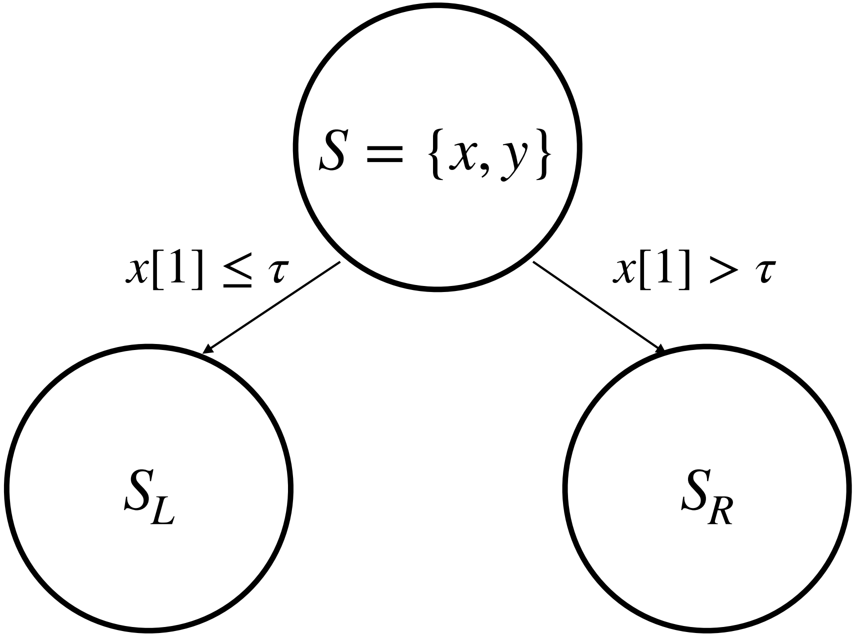
Consider k-class classification, i.e., $y \in \{1, 2, \ldots, k\}$



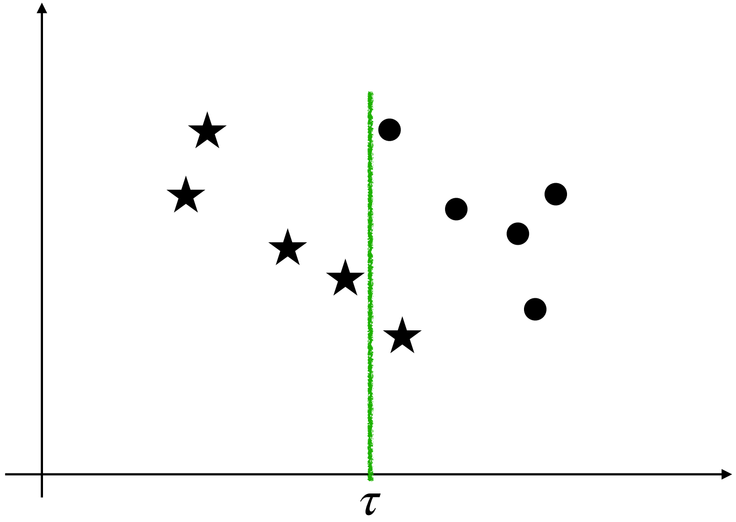


How to split a tree node

Consider k-class classification, i.e., $y \in \{1,2,\ldots,k\}$

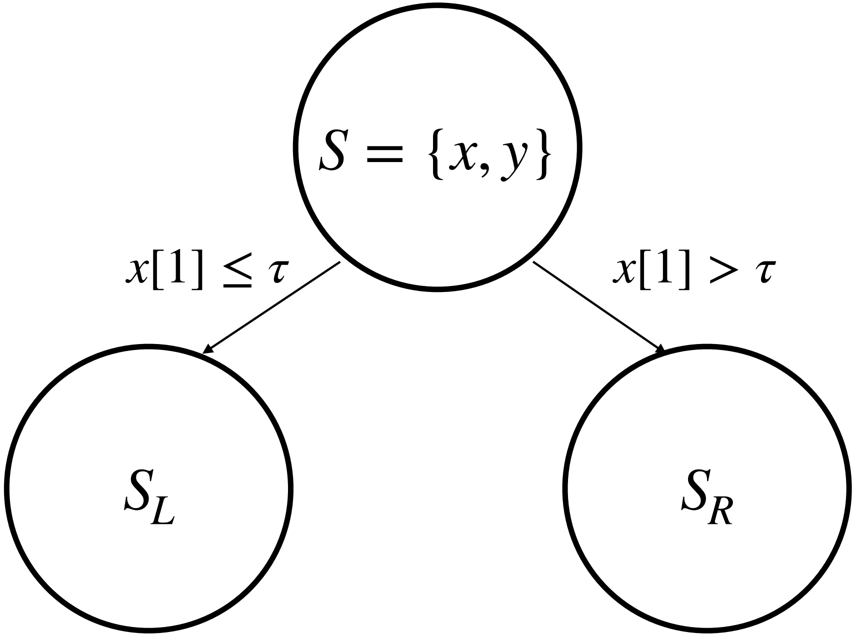


$S_L + S_R = S,\ S_L \cap S_R = \emptyset$



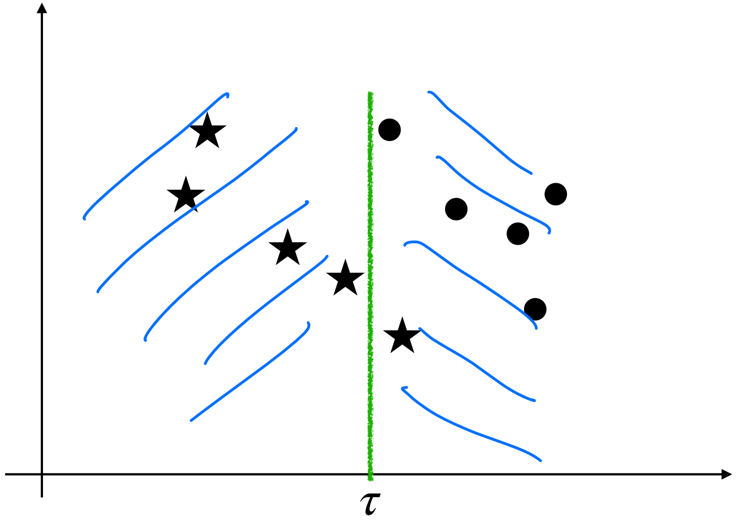
How to split a tree node

Consider k-class classification, i.e., $y \in \{1, 2, \ldots, k\}$



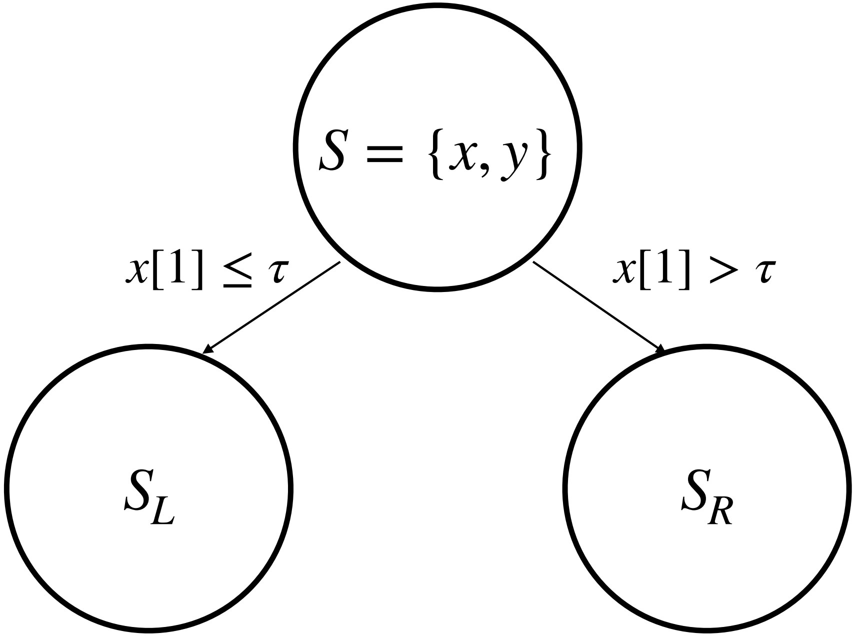
$S_L + S_R = S$, $S_L \cap S_R = \emptyset$

Goal: do an axis aligned split such that diversity of labels in leafs are reduced



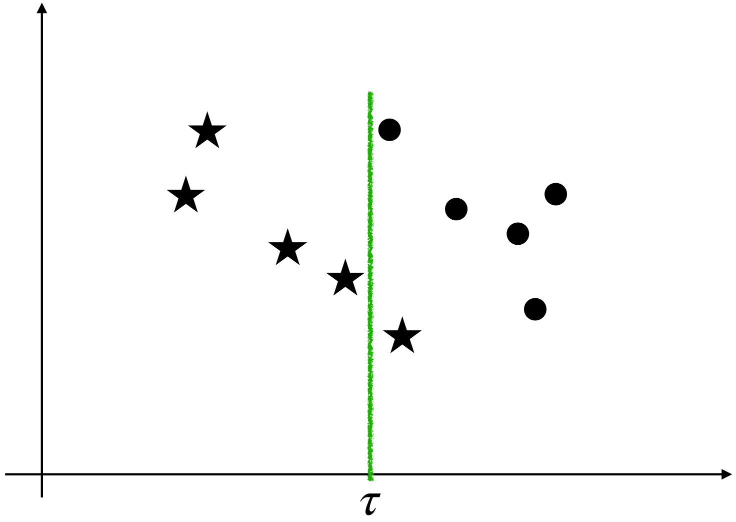
How to split a tree node

Consider k-class classification, i.e., $y \in \{1,2,\ldots,k\}$



$S_L + S_R = S,\ S_L \cap S_R = \emptyset$

Goal: do an axis aligned split such that diversity of labels in leafs are reduced



How to mathematically quantify “diversity”?

Detour: Entropy

Given a set $S = \{x_i, y_i\}_{i=1}^n$, $y_i \in \{1,2,\ldots,k\}$, measure the diversity of labels via entropy

Detour: Entropy

Given a set $S = \{x_i, y_i\}_{i=1}^n$, $y_i \in \{1,2,\ldots,k\}$, measure the diversity of labels via entropy

1. For each label i, Define $p_i = \dfrac{\text{number of label i}}{n} = \dfrac{\sum_{j=1}^n \mathbf{1}(y_j = i)}{n}$

Detour: Entropy

Given a set $S = \{x_i, y_i\}_{i=1}^{n}$, $y_i \in \{1,2,\ldots,k\}$, measure the diversity of labels via entropy

1. For each label i, Define $p_i = \dfrac{\text{number of label i}}{n} = \dfrac{\sum_{j=1}^{n} \mathbf{1}(y_j = i)}{n}$

(Probability of y being label i)

Detour: Entropy

Given a set $S = \{x_i, y_i\}_{i=1}^n$, $y_i \in \{1,2,\ldots,k\}$, measure the diversity of labels via entropy

1. For each label i, Define $p_i = \dfrac{\text{number of label i}}{n} = \dfrac{\sum_{j=1}^n \mathbf{1}(y_j = i)}{n}$

(Probability of y being label i)

2. Entropy: $H(S) = \sum_{i=1}^{k} - p_i \ln(p_i)$

Detour: Entropy

Given a set $S = \{x_i, y_i\}_{i=1}^n$, $y_i \in \{1,2,\ldots,k\}$, measure the diversity of labels via entropy

1. For each label i, Define $p_i = \frac{\text{number of label i}}{n} = \frac{\sum_{j=1}^n \mathbf{1}(y_j = i)}{n}$

(Probability of y being label i)

2. Entropy: $H(S) = \sum_{i=1}^{k} -p_i \ln(p_i)$

High entropy means “diverse, chaos, uncertain”

Entropy

Consider a Bernoulli distribution

$$P(y = 1) = p, P(y = 0) = 1 - p$$

Entropy $H(y)$:

Entropy

Consider a Bernoulli distribution

$$P(y=1)=p,\ P(y=0)=1-p$$

Entropy $H(y)$:

$$-P(y=1)\cdot\ln P(y=1)-P(y=0)\cdot\ln P(y=0)$$

p

1-p

Entropy

Consider a Bernoulli distribution

$$P(y = 1) = p, P(y = 0) = 1 - p$$

Entropy $H(y)$**:**

$$-P(y = 1) \cdot \ln P(y = 1) - P(y = 0) \cdot \ln P(y = 0)$$

$$= -p \ln p - (1 - p)\ln(1 - p)$$

Entropy

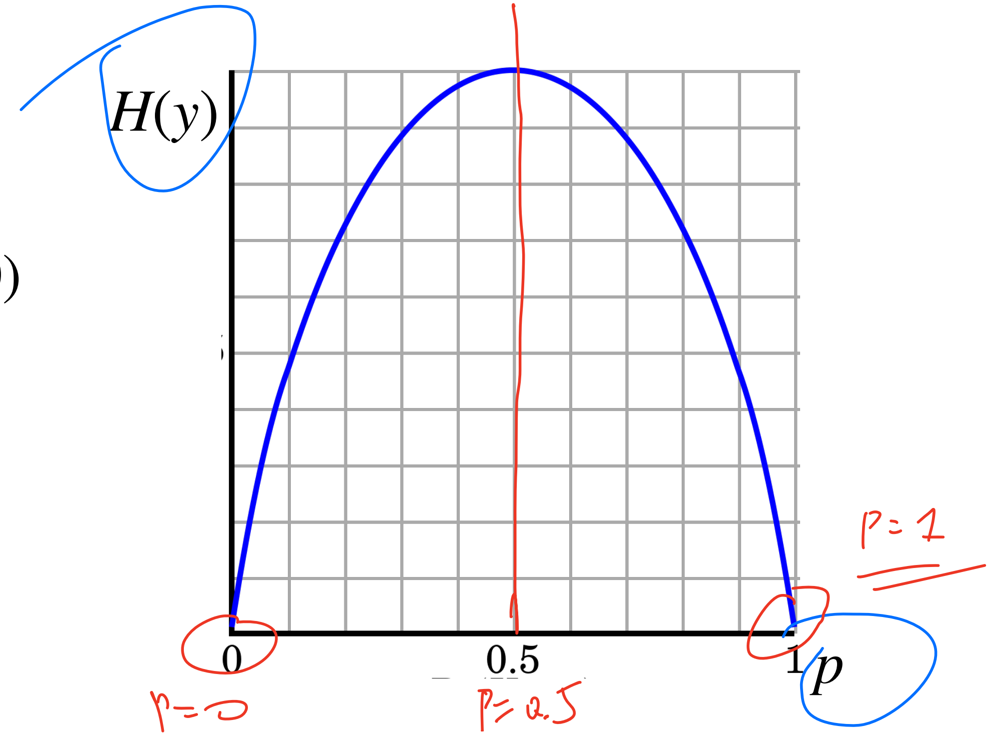
Consider a Bernoulli distribution

$$P(y = 1) = p,\ P(y = 0) = 1 - p$$

Entropy $H(y)$**:**

$$-P(y=1) \cdot \ln P(y=1) - P(y=0) \cdot \ln P(y=0)$$

$$= -p \ln p - (1-p)\ln(1-p)$$



Entropy

Consider a categorical distribution

$$y \in \{1,2,\ldots,k\}, P(y=i) = p_i \geq 0, \sum_{i=1}^{k} p_i = 1$$

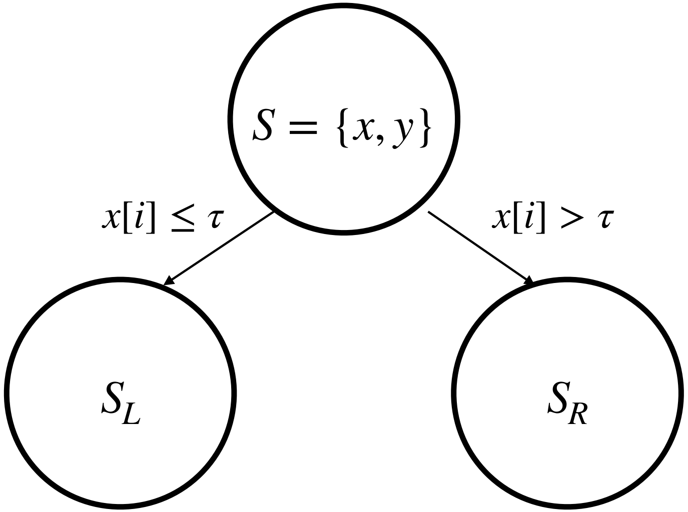
$$\sum_{i=1}^{k} -(p_i \ln p_i)$$

Q: when is entropy maximized?

$$p_1 = p_2 \ldots = p_k = \frac{1}{k}$$

Back to tree node split…

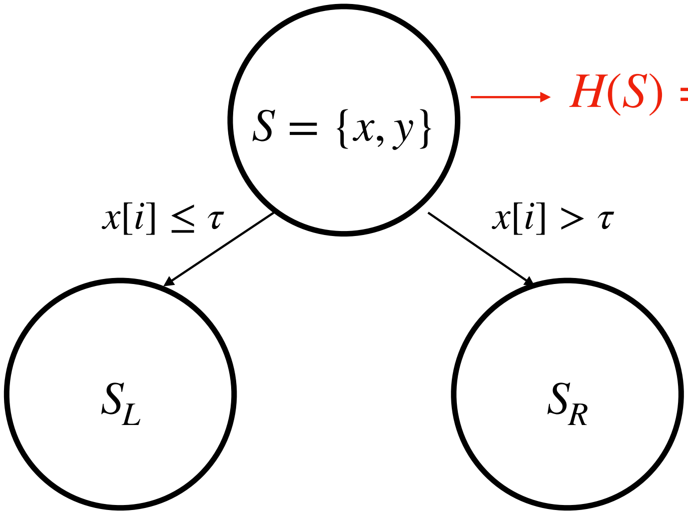
Consider a split, i.e, dim i and threshold τ,



Optimization:

Back to tree node split…

Consider a split, i.e, dim i and threshold τ,



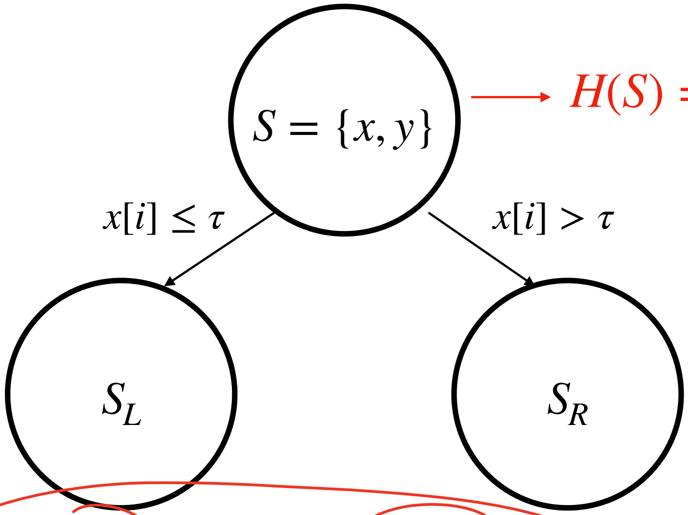
$$H(S) = \sum_{i=1}^{k} - p_i \ln p_i$$

$$p_i = \frac{\#\text{ of label } i \text{ in } S}{|S|}$$

Optimization:

Back to tree node split…

Consider a split, i.e, dim i and threshold τ,



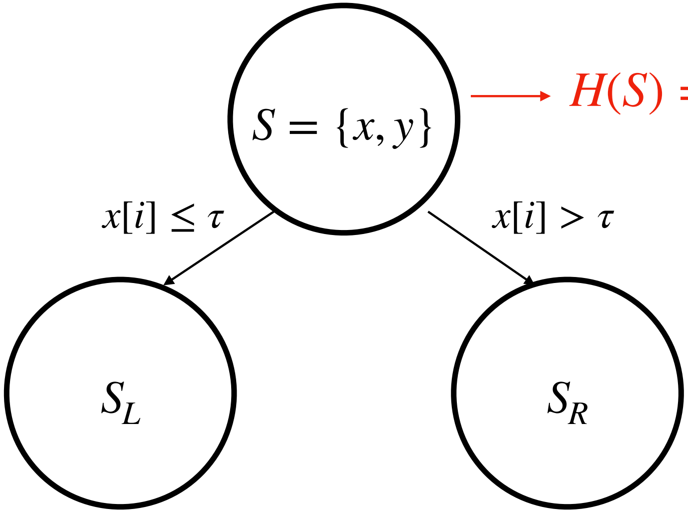
$$H(S) = \sum_{i=1}^{k} -p_i \ln p_i$$

$$\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$$

Optimization:

Back to tree node split…

Consider a split, i.e, dim i and threshold τ,



$$H(S) = \sum_{i=1}^{k} - p_i \ln p_i$$

$$\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$$

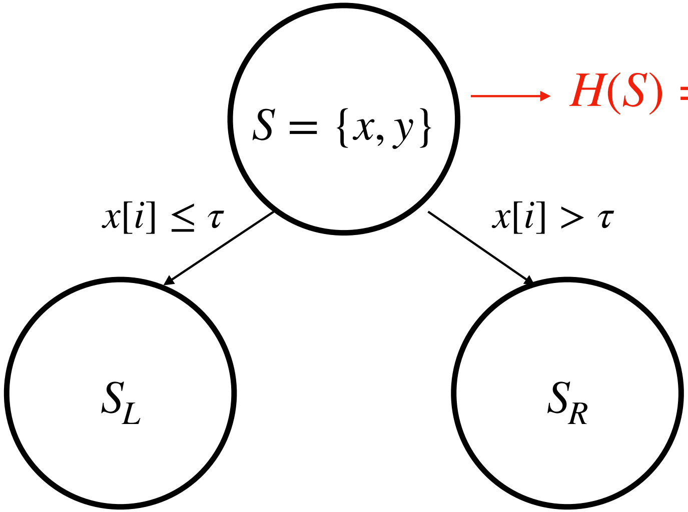
Optimization:

Find a split (i, τ) such that

$\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$ is the smallest

Back to tree node split…

Consider a split, i.e, dim i and threshold τ,



$$H(S) = \sum_{i=1}^{k} -p_i \ln p_i$$

$$\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$$

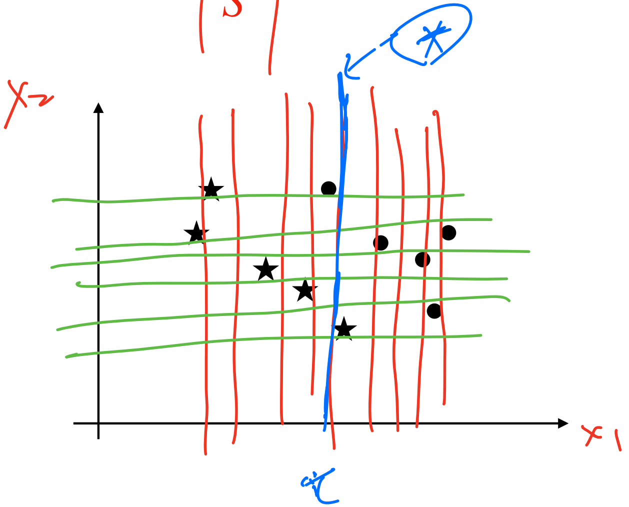
Optimization:

dim from 1,2,…,d

$\tau \in \mathbb{R}$ threshold

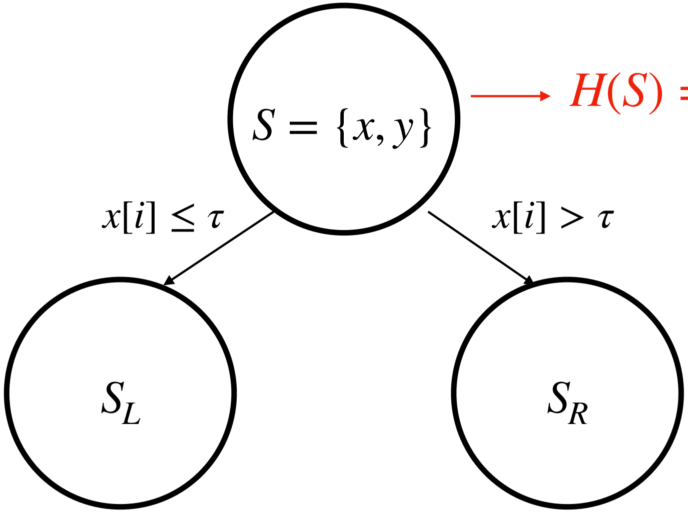
Find a split (i, τ) such that

$\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$ is the smallest



Back to tree node split…

Consider a split, i.e, dim i and threshold τ,



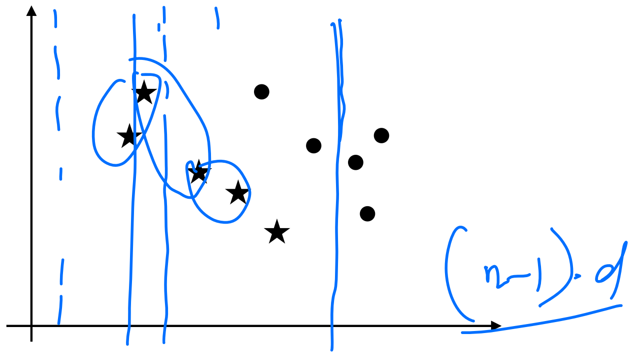
$$H(S) = \sum_{i=1}^{k} -p_i \ln p_i$$

$$\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$$

Optimization:

Find a split (i, τ) such that

$$\frac{|S_L|}{|S|} H(S_L) + \frac{|S_R|}{|S|} H(S_R)$$ is the smallest



Q: how many splits we need to check?

n points

Put everything together — ID3 algorithm

Input: training set $S = \{x, y\}$

Decision_tree(S):

Put everything together — ID3 algorithm

Input: training set $S = \{x, y\}$

Decision_tree(S):

- If all y in S are the same

Put everything together — ID3 algorithm

Input: training set $S = \{x, y\}$

Decision_tree(S):

- If all y in S are the same

  Done, and return this label

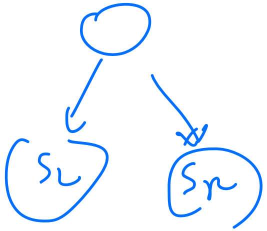
Put everything together — ID3 algorithm

Input: training set $S = \{x, y\}$

Decision_tree(S):

- If all y in S are the same

  Done, and return this label



- Else:

  Find a split (i, τ) that minimizes weighted entropy $\longrightarrow$ $\frac{|S_L|}{|S|} \cdot H(S_L) + \frac{|S_R|}{|S|} \cdot H(S_R)$

Put everything together — ID3 algorithm

Input: training set $S = \{x, y\}$

Decision_tree(S):

- If all y in S are the same

  Done, and return this label

- Else:

  Find a split (i, τ) that minimizes weighted entropy

  Call **Decision_tree(S_L)** & **Decision_tree(S_R)**

Put everything together — ID3 algorithm

Input: training set $S = \{x, y\}$

Decision_tree(S):

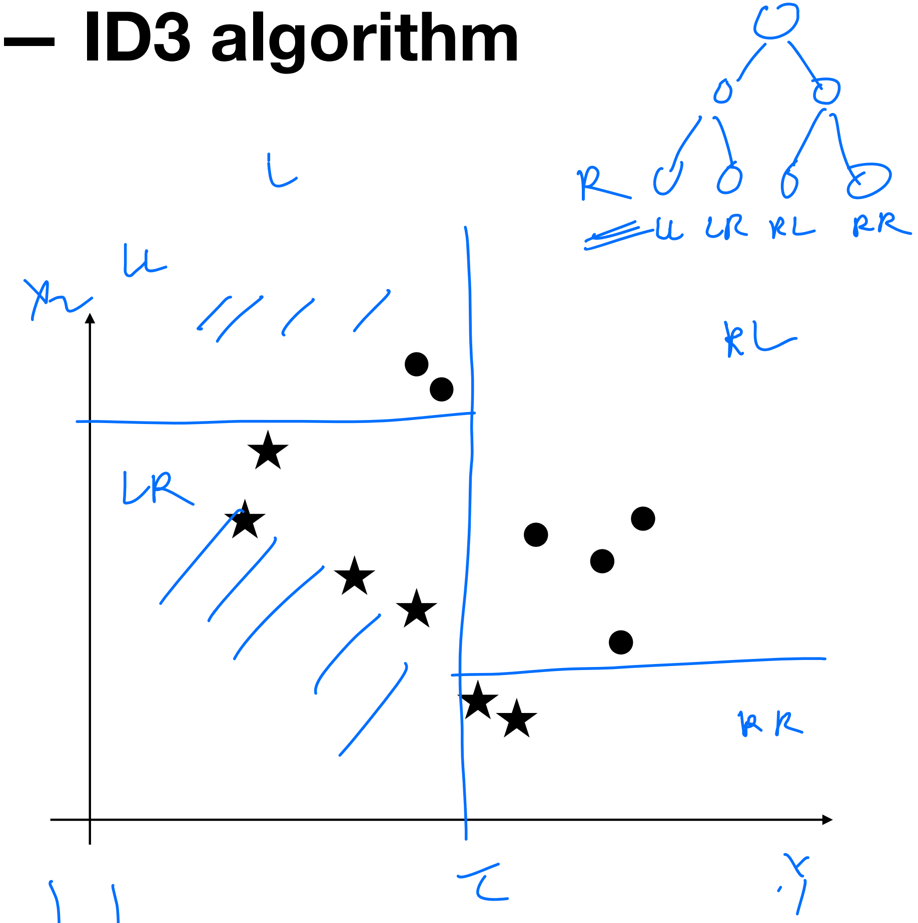
- If all y in S are the same

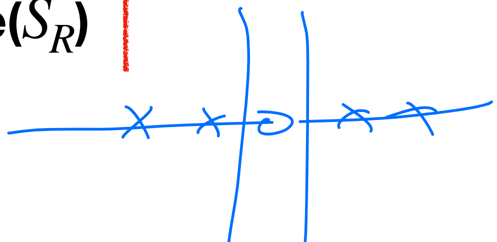
  Done, and return this label

- Else:

  Find a split (i, τ) that minimizes weighted entropy

  Call **Decision_tree(S_L)** & **Decision_tree(S_R)**





Outline of Today

1. Decision tree in classification

2. Decision tree in regression

3. Demos of decision tree

Regression

How to split the note, i.e., what is the diversity measure?

$y \in \mathbb{R}$

Regression

How to split the note, i.e., what is the diversity measure?

Consider a set of training points $S = \{x_i, y_i\}_{i=1}^m, y_i \in \mathbb{R}$

Regression

How to split the note, i.e., what is the diversity measure?

Consider a set of training points $S = \{x_i, y_i\}_{i=1}^{m}, y_i \in \mathbb{R}$

Define the sample mean $\hat{y}_S = \sum_{i=1}^{m} y_i / m$

Regression

How to split the note, i.e., what is the diversity measure?

Consider a set of training points $S = \{x_i, y_i\}_{i=1}^m, y_i \in \mathbb{R}$

Define the sample mean $\hat{y}_S = \sum_{i=1}^m y_i / m$

Impurity: sample variance $\widehat{Var}(S) = \sum_{i=1}^m (y_i - \bar{y}_S)^2 / m$

$\Rightarrow Var(y)$ when $m \to \infty$

Regression Tree

Regression_Tree(S):

Regression Tree

Threshold $k \in N^+$

(e.g. $k = 3$)

Regression_Tree(S):

- IF $|S| \leq k$:

  Set leaf value to be $\bar{y}_S$

Regression Tree

Regression_Tree(S):

- IF $|S| \leq k$:

  Set leaf value to be $\bar{y}_S$

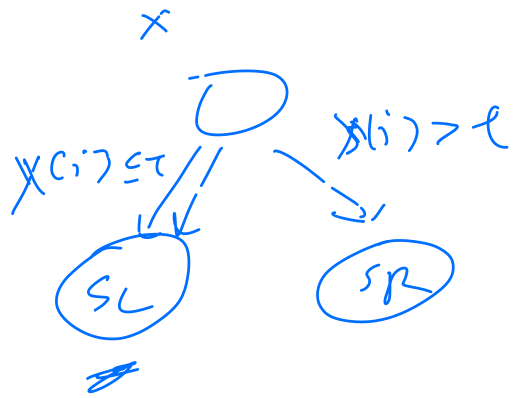
- ELSE:

Regression Tree

Regression_Tree(S):

- IF $|S| \leq k$:
  Set leaf value to be $\bar{y}_S$
- ELSE:
  For all (i, τ), find the split such that minimizes $\frac{|S_L|}{|S|}\widehat{Var}(S_L) + \frac{|S_R|}{|S|}\widehat{Var}(S_R)$

Regression Tree



Regression_Tree(S):

- IF $S \leq k$:

  Set leaf value to be $\bar{y}_S$

- ELSE:

  For all (i, τ), find the split such that minimizes $\frac{S_L}{S}\widehat{Var}(S_L) + \frac{S_R}{S}\widehat{Var}(S_R)$

  Call Regression_Tree(S_L) & Regression_Tree(S_R)

Outline of Today

1. Decision tree in classification

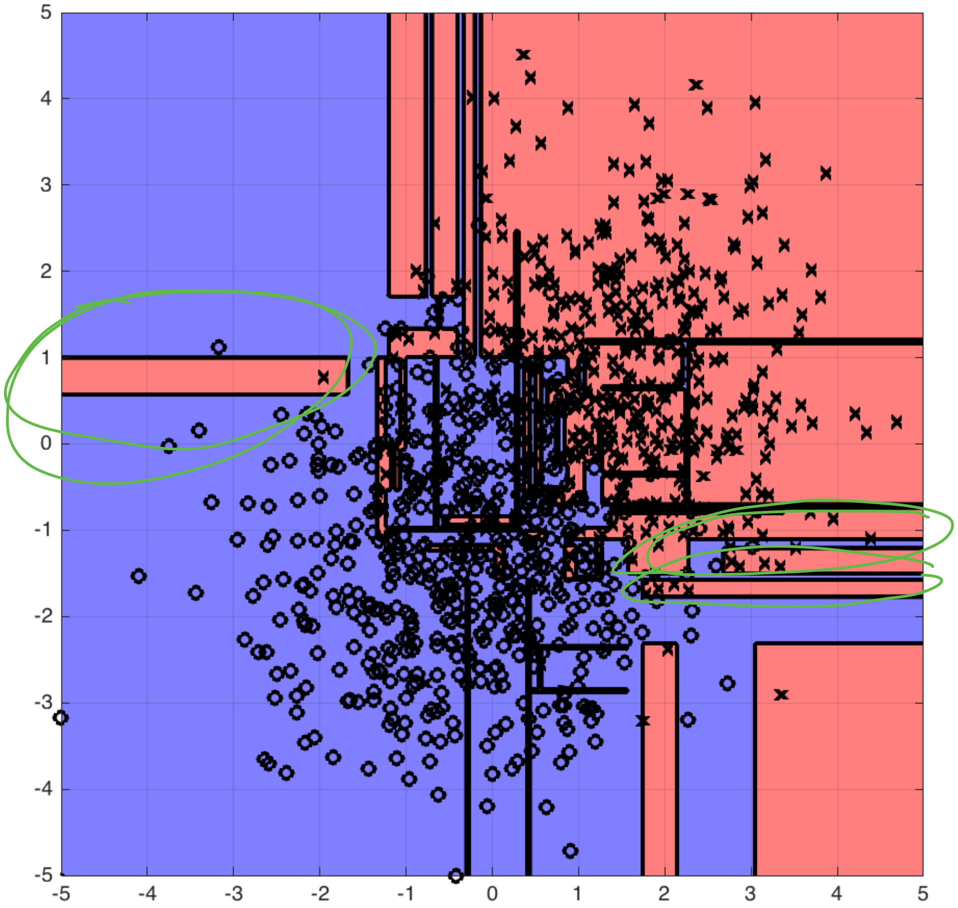
2. Decision tree in regression

3. Demos of decision tree

Issue of Decision Trees

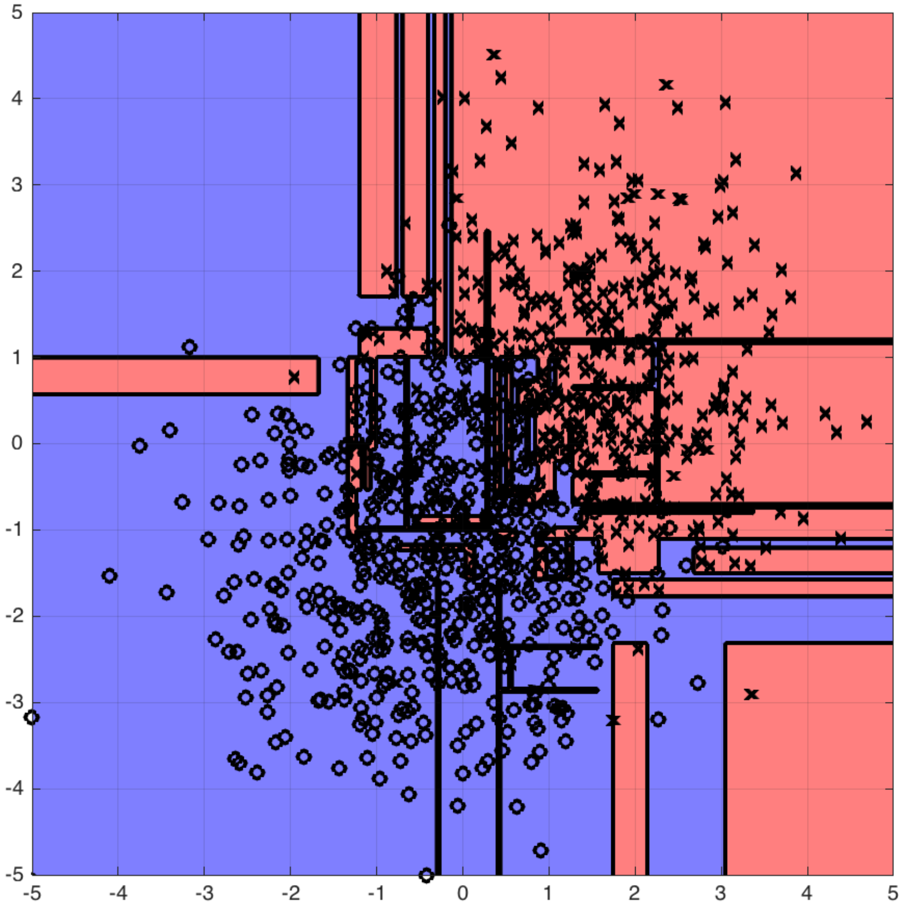
Decision Tree can have high variance, i.e., overfilling!

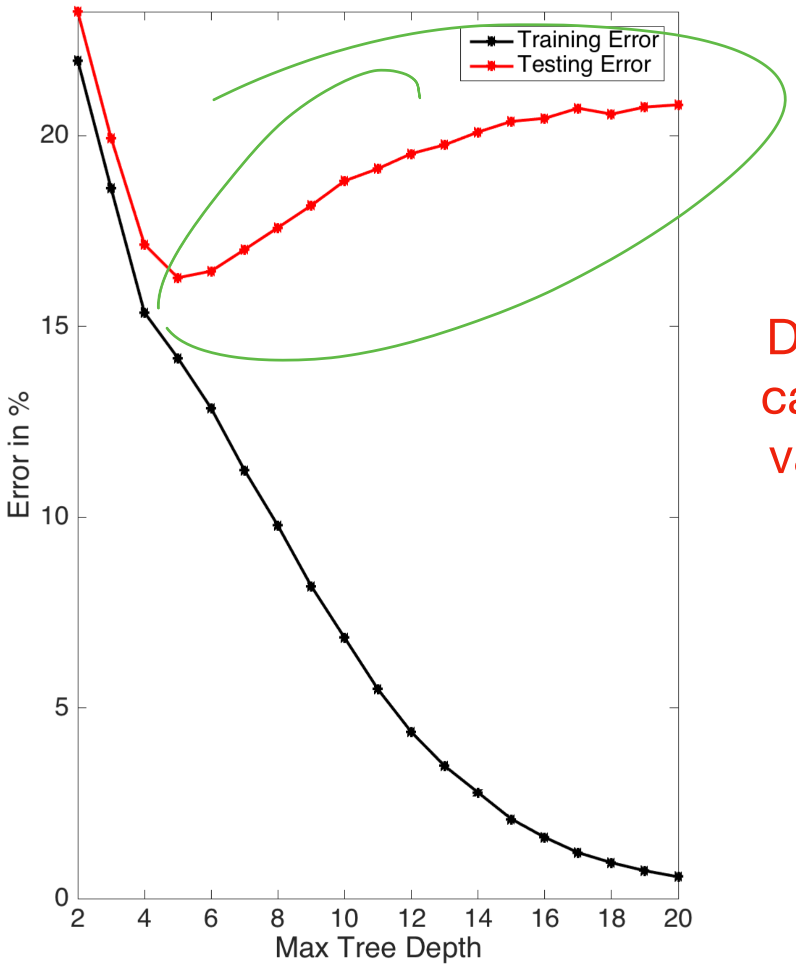
Issue of Decision Trees



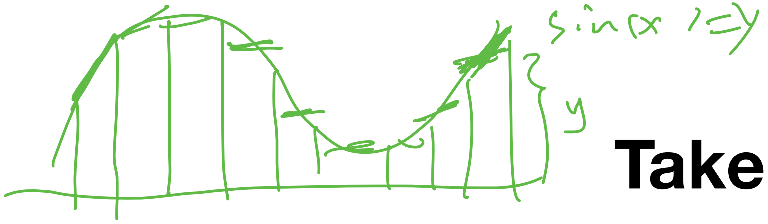
Decision Tree can have high variance, i.e., overfilling!

Issue of Decision Trees





Decision Tree can have high variance, i.e., overfilling!



Take-home messages

1 Decision tree algorithms splits space into axis-aligned regions

Each region ideally should only contain one unique label

2: Split a node such that the entropy of labels in the leafs are minimized

3: Can easily overfit as the depth of the tree increases
(limiting the depth of the tree is a good regularization)