



CIS 678 - Machine Learning

ML Models: Decision Tree



Decision Tree

- Another non-parametric model
 - Recall k-NN, its an in memory model; right?



Decision Tree

- Another non-parametric model
 - Recall k-NN, its an in memory model; right?
- Decision Tree is our second example

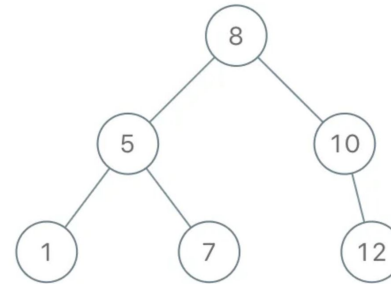


Decision Tree

- Another non-parametric model
 - Recall k-NN, its an in memory model; right?
- Decision Tree is our second example
 - Those are with CS background are already aware of BST
 - Whiteboarding

Decision Tree

- Another non-parametric model
 - Recall k-NN, its an in memory model; right?
- Decision Tree is our second example
 - Those are with CS background are already aware of **BST**
 - Whiteboarding



8, 5, 10, 1, 12, 7



Decision Tree

- *Concepts and Principles*
- *Let's learn through an example*

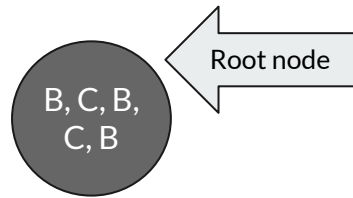


Decision Tree

- Data records for two animal classes:
 - Bunny and Cat

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

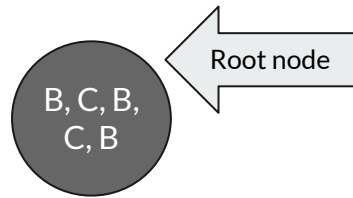
Decision Tree



- What feature should we use to split records?

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

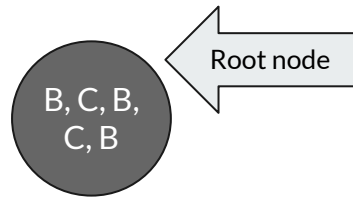
Decision Tree



- What feature should we use to split records?
- nb of legs is useless as there is no variation.

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

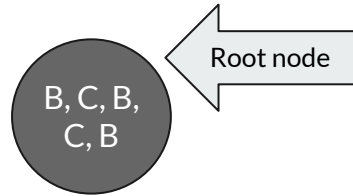
Decision Tree



- What feature should we use to split records?
- nb of legs is useless as there is no variation.
- We can use the **'weight(lb)'** feature.

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

Decision Tree

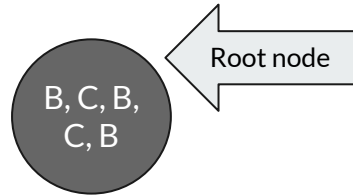


- *Let's plot the data points*

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny



Decision Tree



- *Let's plot the data points*

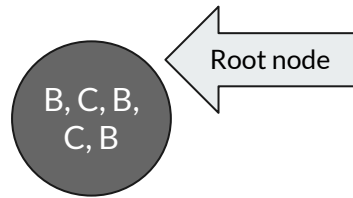


A table with three columns: "nb of legs", "weight (lb)", and "animal". It contains five rows of data. A red dashed rectangle highlights the entire table. A green dashed rectangle highlights the "weight (lb)" column.

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny



Decision Tree

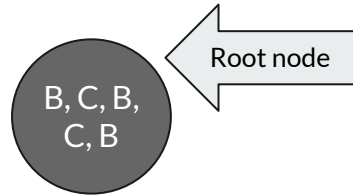


- *Let's plot the data points*

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny



Decision Tree

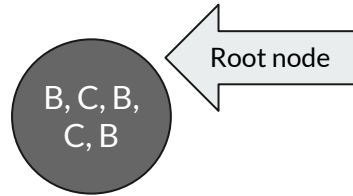


- *Let's plot the data points*

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny



Decision Tree

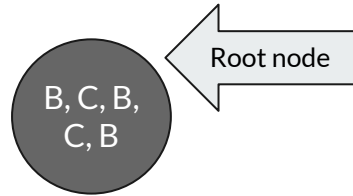


- *Let's plot the data points*

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny



Decision Tree

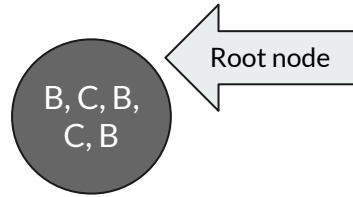


- *Let's plot the data points*

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny



Decision Tree

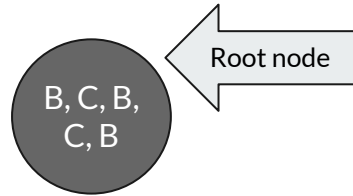


- *Can we identify groups?*

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

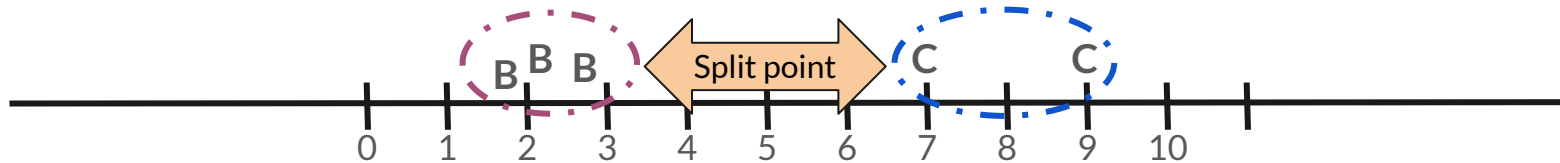


Decision Tree

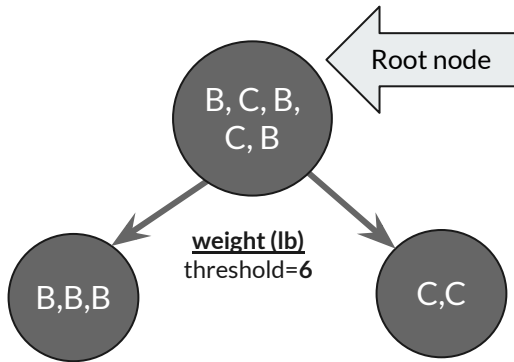


- *Let's find a split point.*

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

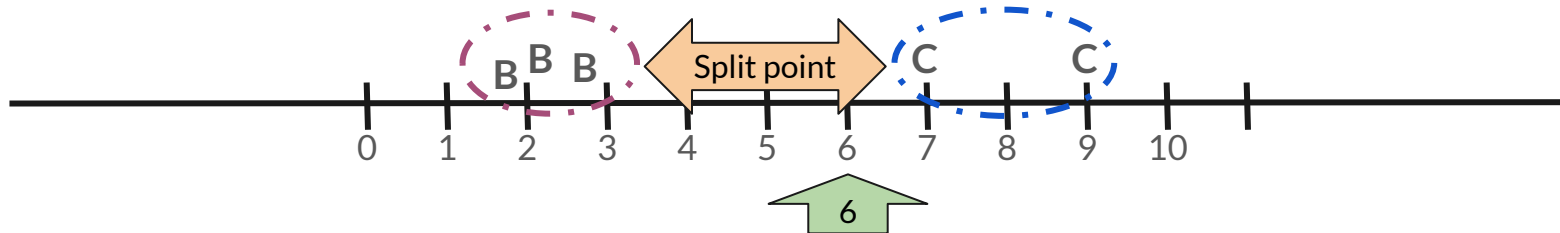


Decision Tree

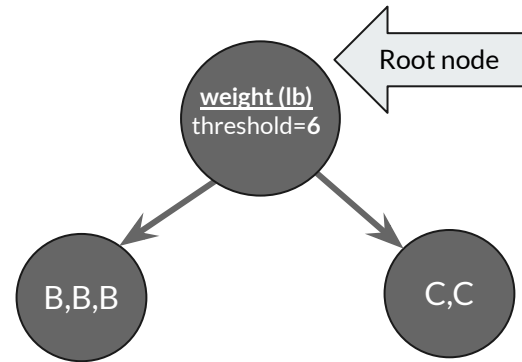


- Create branches

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

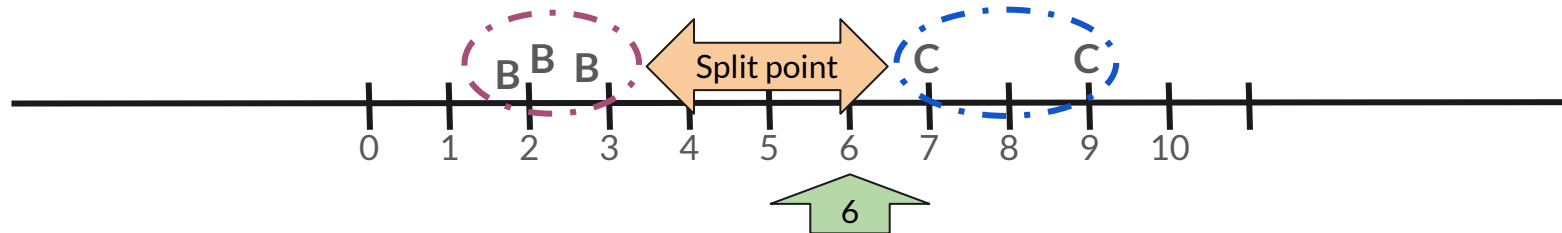


Decision Tree

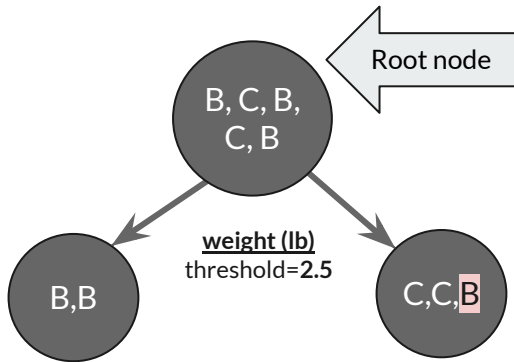


- Create branches

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

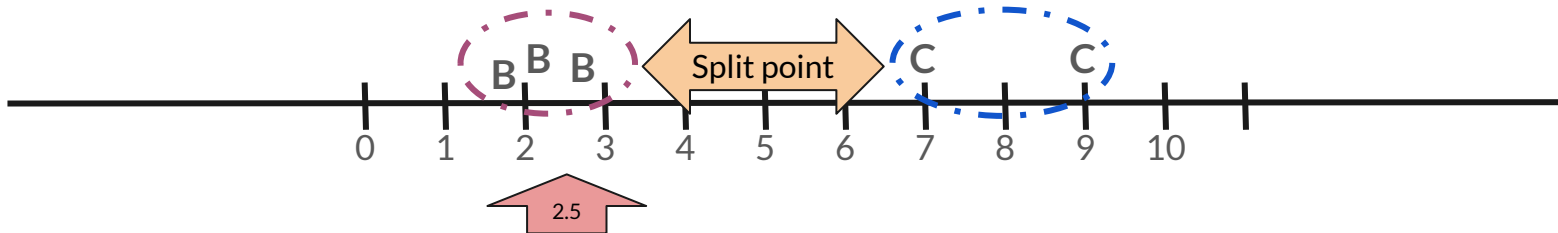


Decision Tree

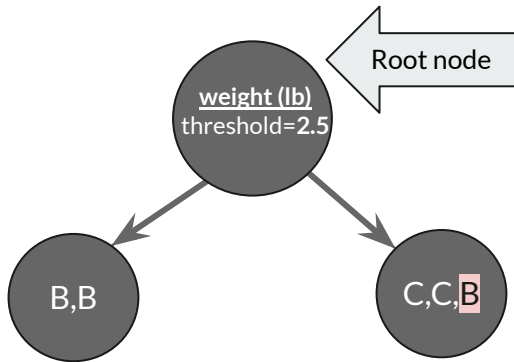


nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

- What if we used a threshold=2.5?

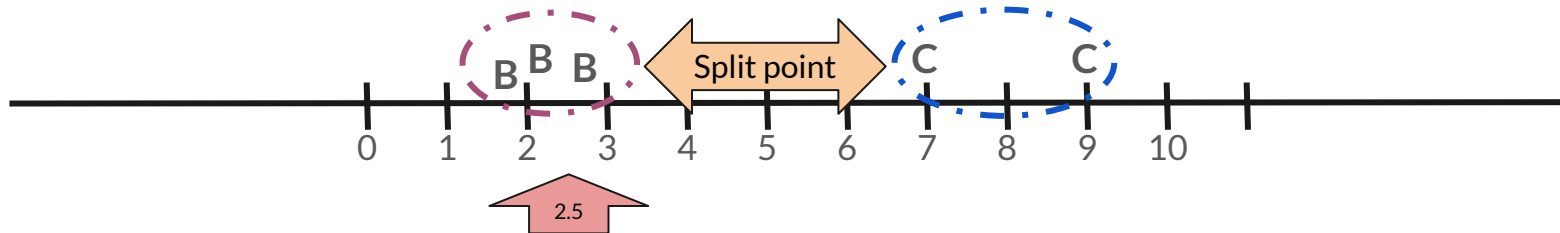


Decision Tree



nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny

- What if we used a threshold=2.5?





Decision Tree

- What metrics are used for split point determination?



Decision Tree

- What metrics are used for split point determination?

Entropy

Gini Impurity

Information Gain

The idea is quite simple, choose the one that make classes more separable.



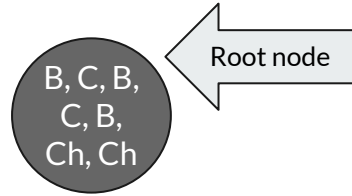
Decision Tree

- How about this configuration?
- We have data points for an additional animal class “Chicken”

nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny
2	2.5	Chicken
2	3	Chicken

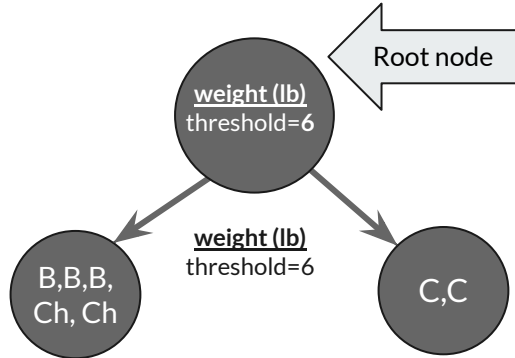


Decision Tree



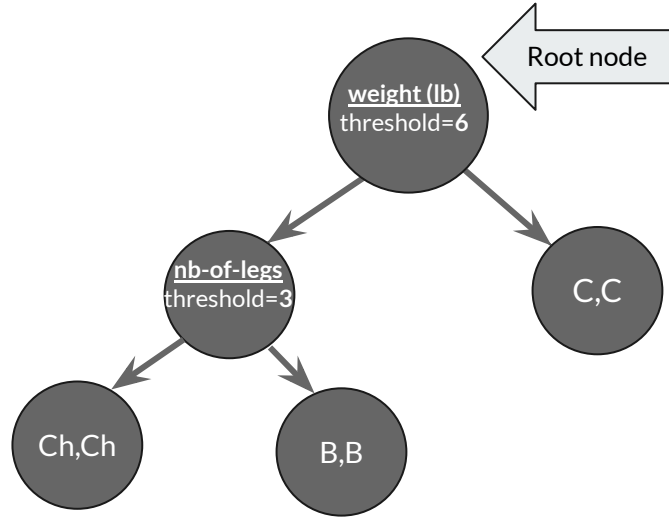
nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny
2	2.5	Chicken
2	3	Chicken

Decision Tree



nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny
2	2.5	Chicken
2	3	Chicken

Decision Tree



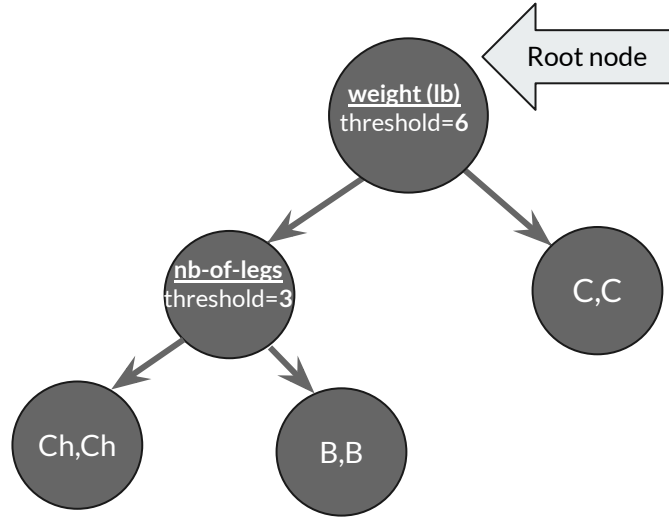
nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny
2	2.5	Chicken
2	3	Chicken



Decision Tree

- Prediction Time

Decision Tree

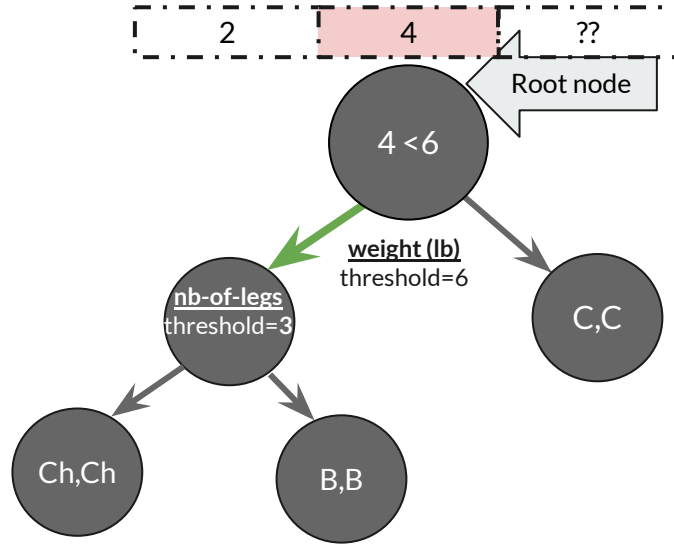


nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny
2	2.5	Chicken
2	3	Chicken

Test case

2	4	??
---	---	----

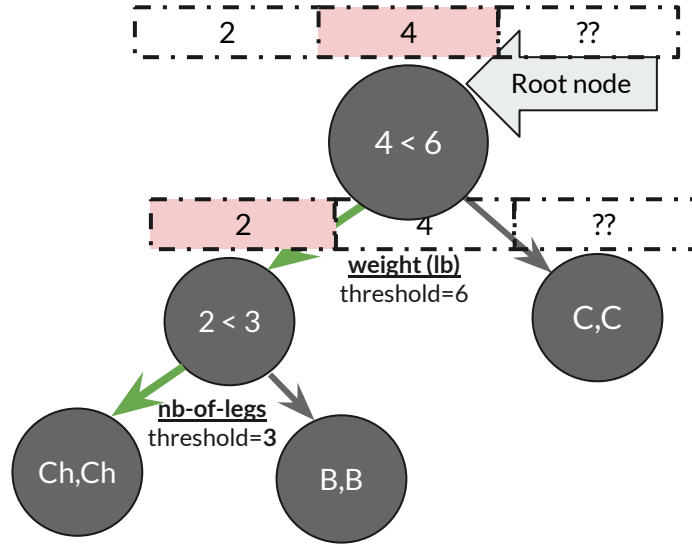
Decision Tree



Test case

2	4	??
---	---	----

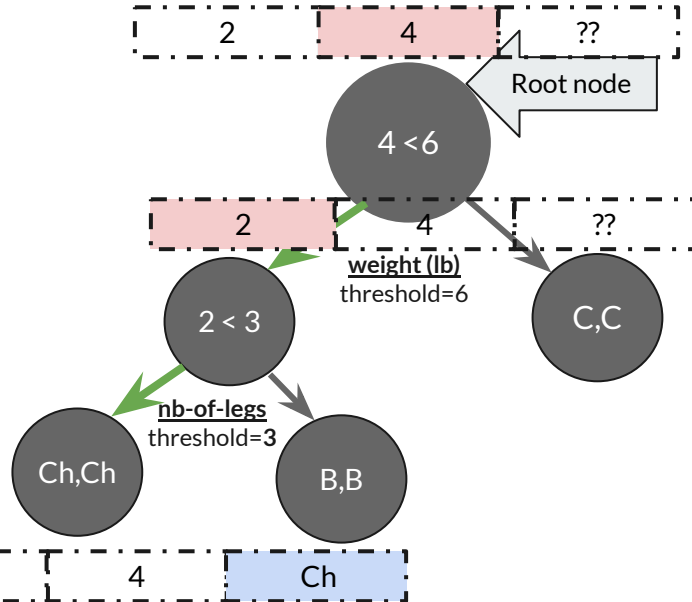
Decision Tree



Test case

2	4	??
---	---	----

Decision Tree



Test case

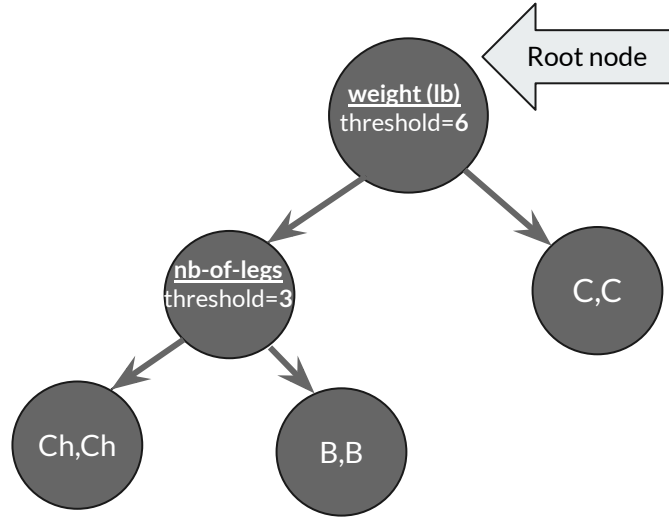
2	4	??
---	---	----



Decision Tree

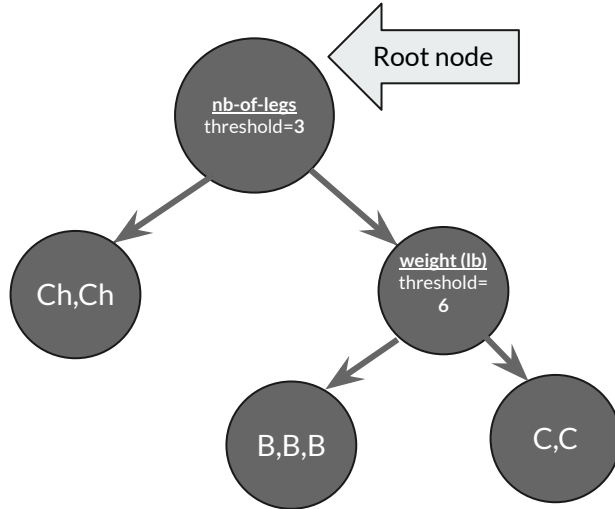
You may have multiple trees

Decision Tree



nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny
2	2.5	Chicken
2	3	Chicken

Decision Tree



nb of legs	weight (lb)	animal
4	2.1	Bunny
4	7	Cat
4	1.7	Bunny
4	9	Cat
4	2.75	Bunny
2	2.5	Chicken
2	3	Chicken



Decision Tree

Entropy and Information Gain



Decision Tree

- What metrics are used for split point determination?

Entropy

Entropy (discrete variable):

$$H(X) = - \sum_{i=1}^n p(x_i) \log_b p(x_i)$$



Decision Tree

- What metrics are used for split point determination?

Entropy

$$IG(T, a) = H(T) - H(T|a),$$

where $H(T|a)$ is the **conditional entropy** of T given the value of **attribute** a .



Decision Tree

How to randomize Trees?



Decision Tree

How to randomize Trees?

Through

- Random sampling of data points
- Random sampling of features
- Randomizing feature combinations



Decision Tree

How to randomize Trees?

Through

- Random sampling of data points
- Random sampling of features
- Randomizing feature combinations

Essentially we can generate many trees for a dataset.



QA