topic 29: decision trees

agenda:

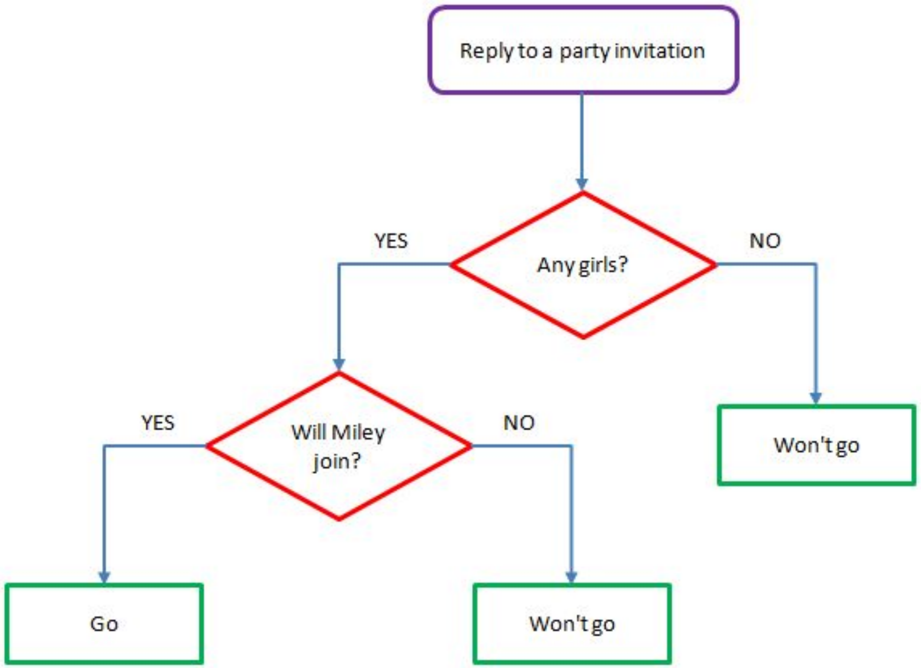
1. what are decision trees?

2. classification trees
    a. how to we grow trees? (cost functions)
    b. implementation
    c. pruning

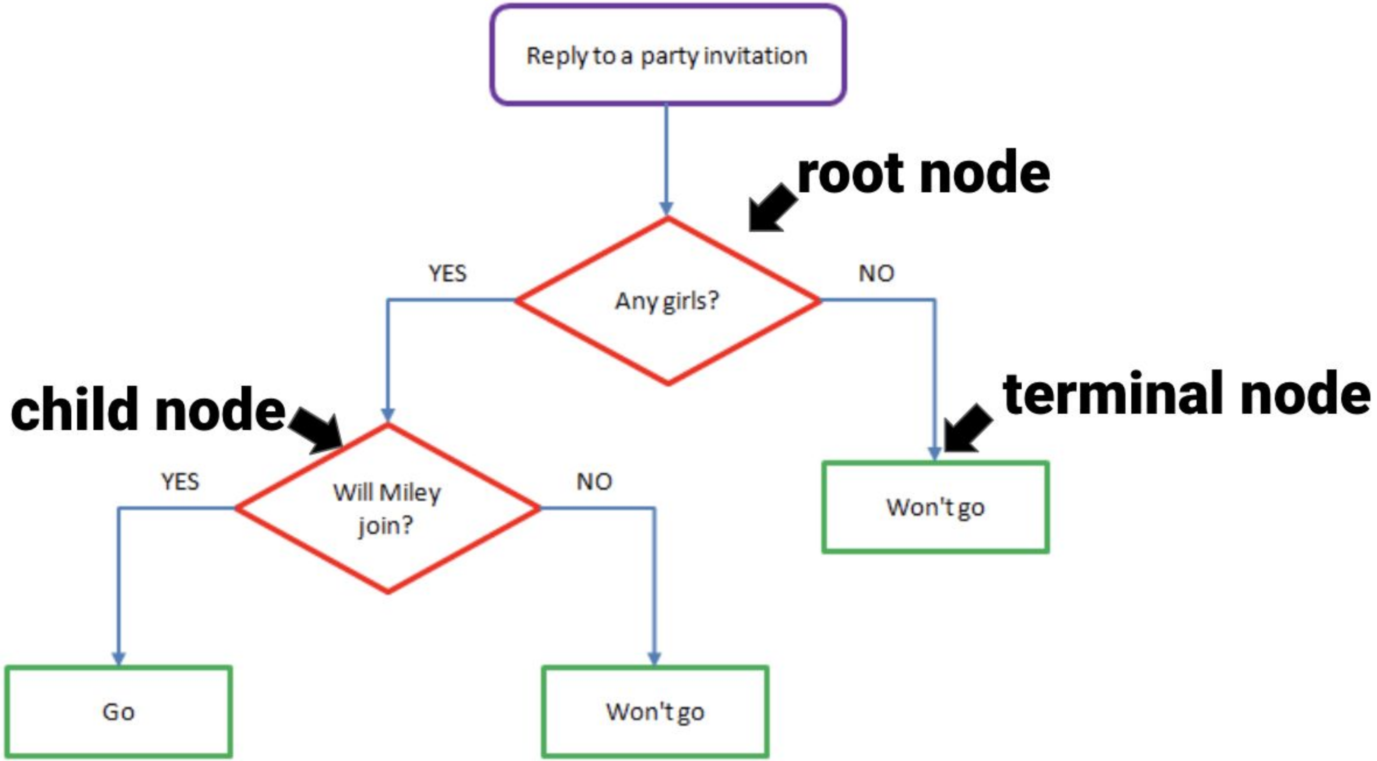
3. regression trees

4. pros and cons of CART trees

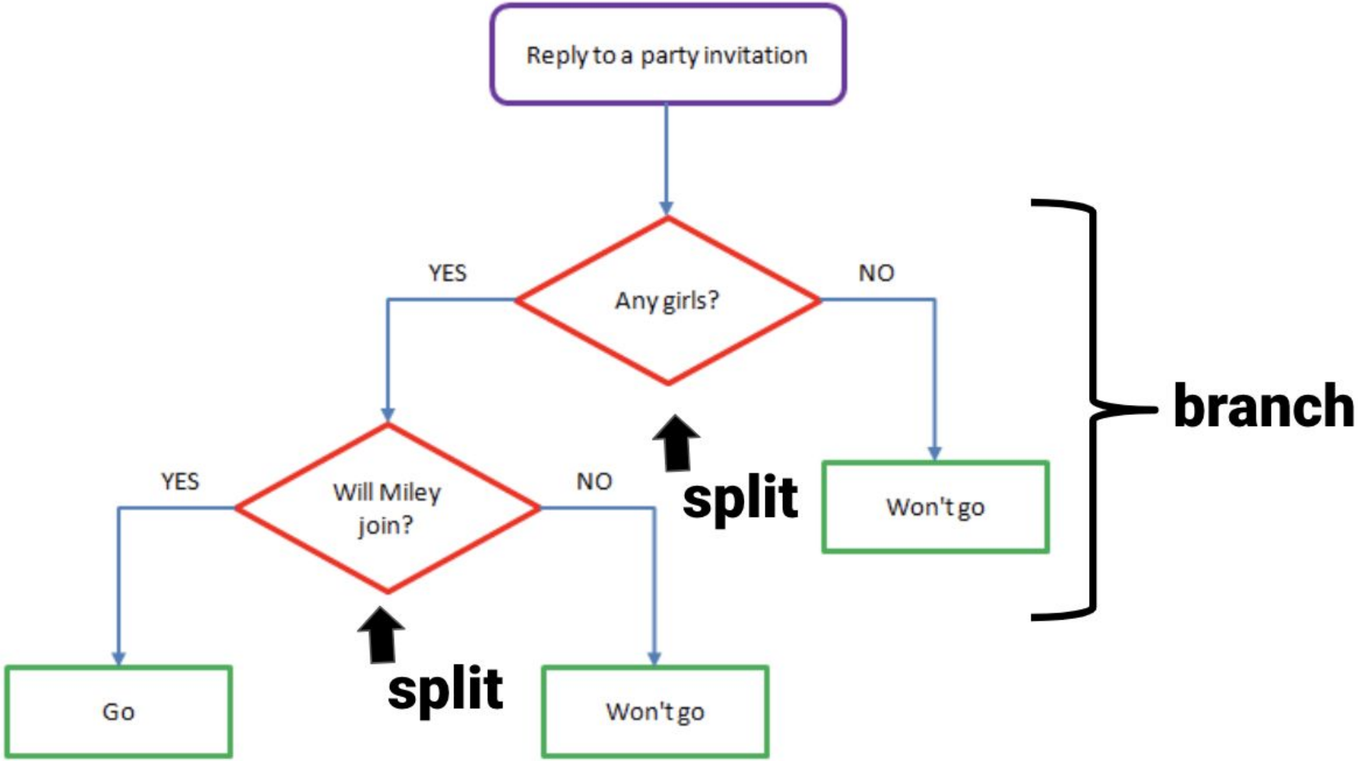
1. what are decision trees?



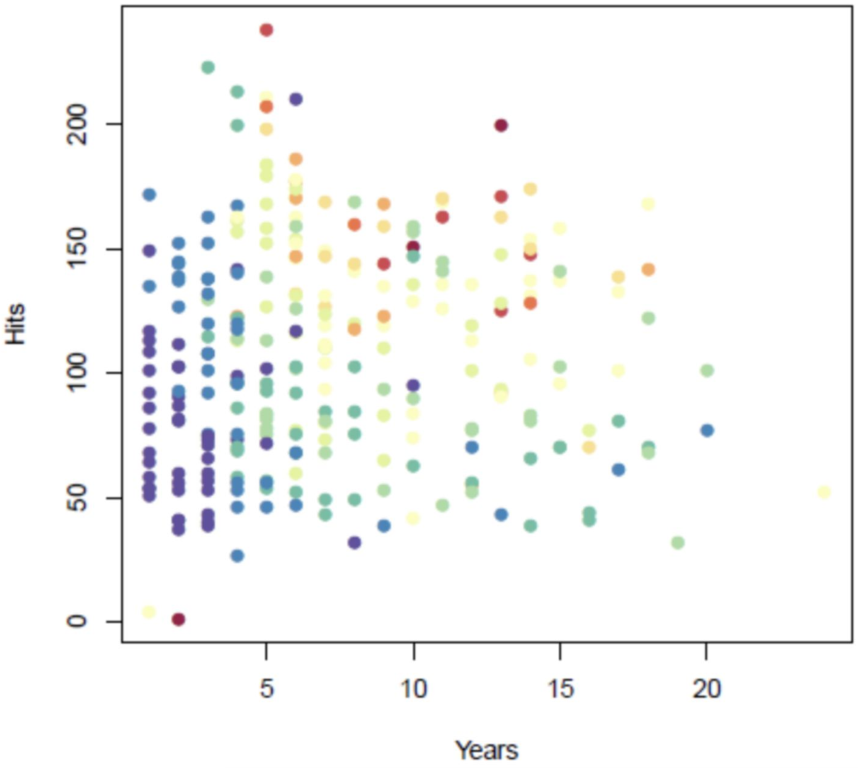
1. what are decision trees? terminology



1. what are decision trees? terminology



2. regression trees

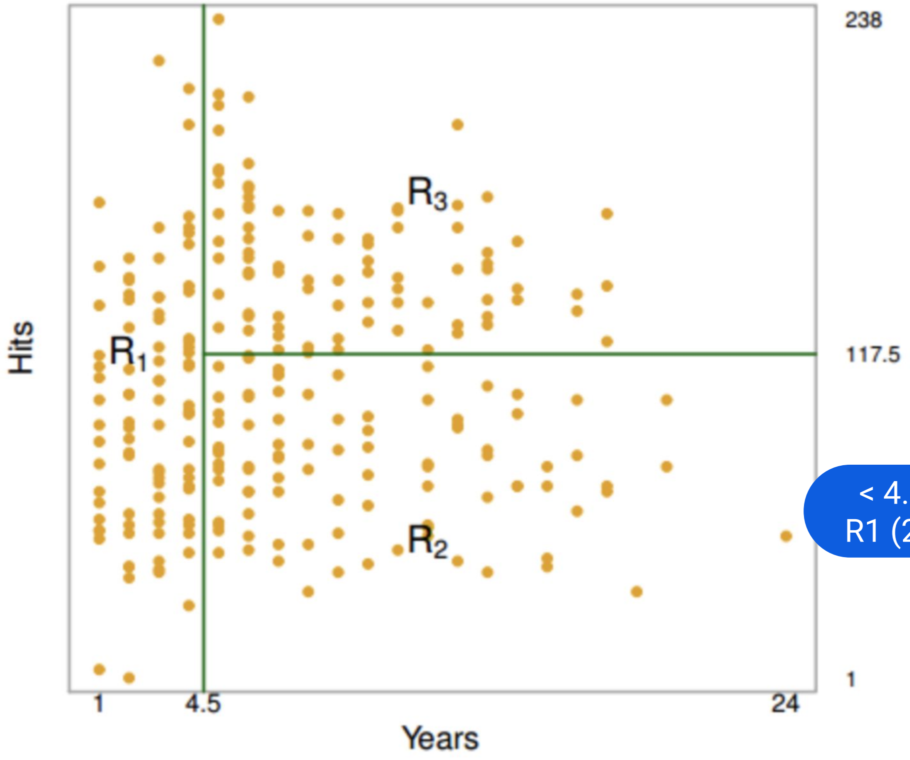


- 2-feature dataset describing professional baseball players

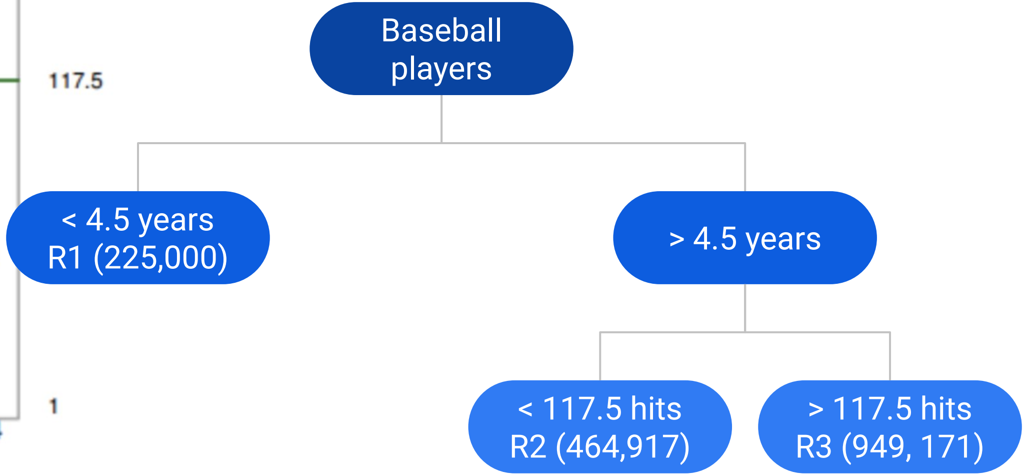
- x-axis: number of years played
- y-axis: number of hits
- color: salary (purple = low, red = high)

- we want to predict **salary**

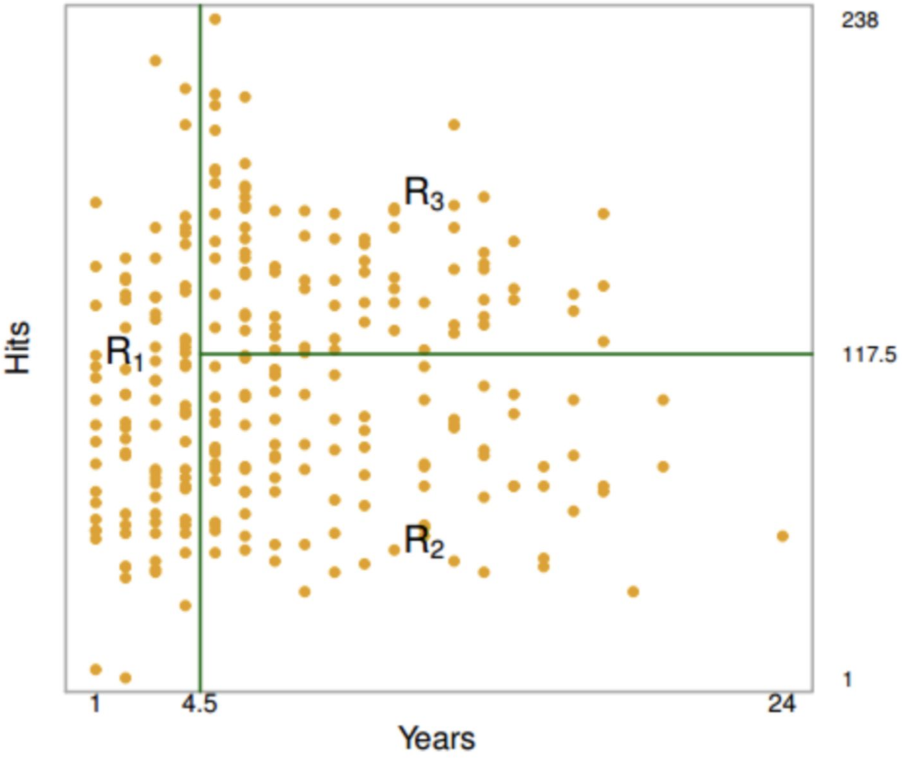
2. building regression trees



- divide the space into **distinct** and **non-overlapping regions**
- the values of the prediction (terminal nodes) are the **mean** value of each region



2. regression trees: cost function



- the **cost function** used to train regression trees is the **MSE**

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$$

- the best split is decided where the overall MSE is lowest

$$R_1(j,s) = \{X | X_j < s\} \text{ and } R_2(j,s) = \{X | X_j \geq s\},$$

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

this approach is **top-down** and **greedy**

2. regression trees: implementation & pruning

to the jupyter notebook!

post-pruning questions:

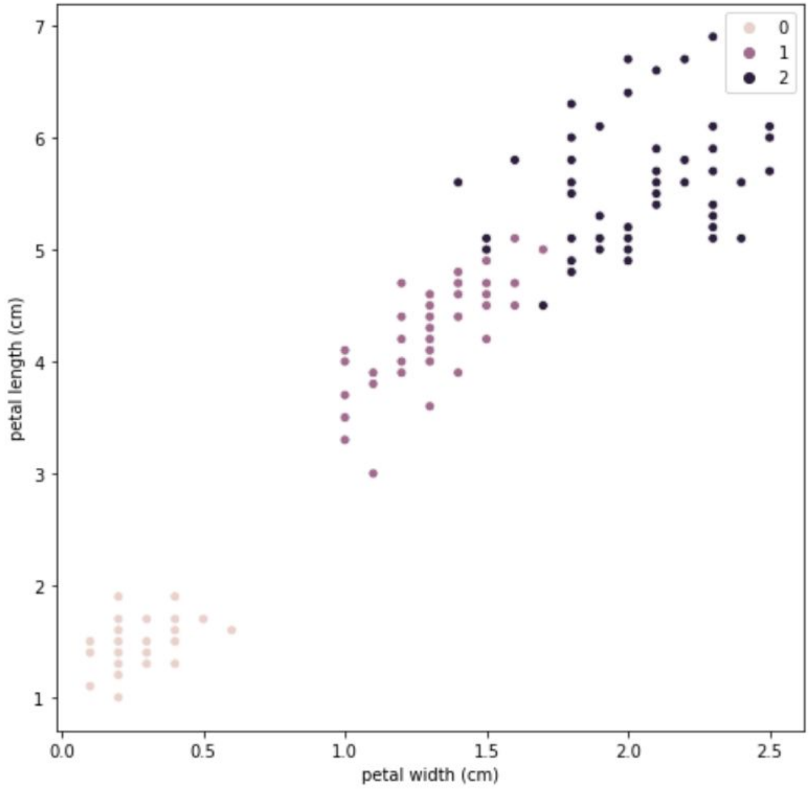
- why don’t we want to make too many splits?
- why a regression tree over a multiple linear regression?

3. classification trees

- what's the difference between classification and regression?

- how are "best-splits" determined?
- what is a hyperparameter we tune for regression trees?

3. classification trees

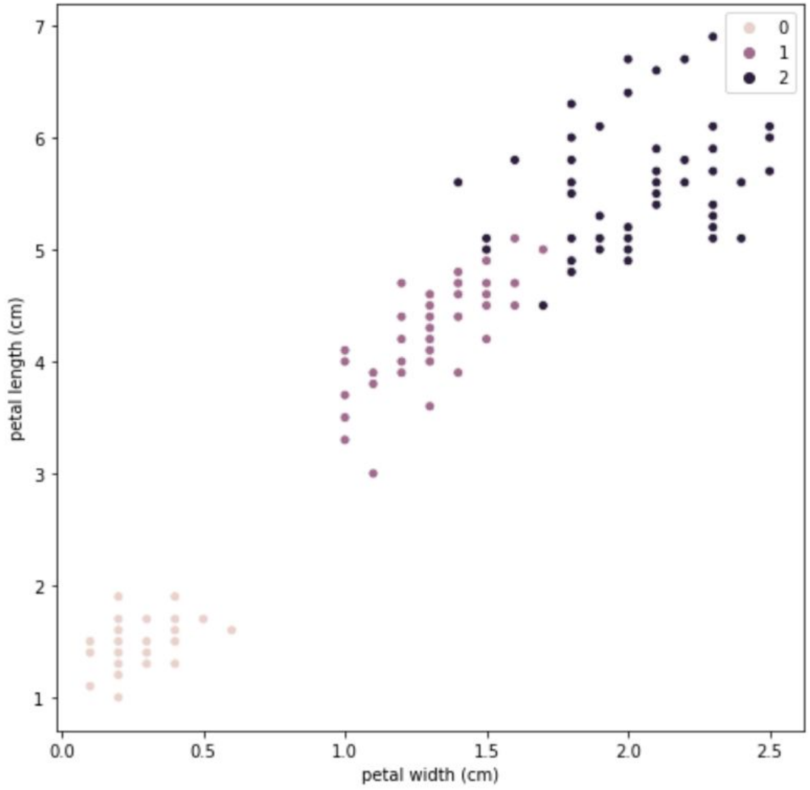


- 2-feature dataset describing irises

- x-axis: petal width
- y-axis: petal length
- color: 3 shades of purple for 3 species

- we want to predict **species**

3. building classification trees



- instead of using MSE (for continuous)
- Gini Purity Index (0-0.5)

$$1 - \sum_{t=0}^{t=k} P_t^2$$

$$1 - \left(\frac{iPhone}{Total}\right)^2 - \left(\frac{Android}{Total}\right)^2 = 1 - \left(\frac{10}{25}\right)^2 - \left(\frac{15}{25}\right)^2 = 0.48$$

- Entropy (0-1)

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

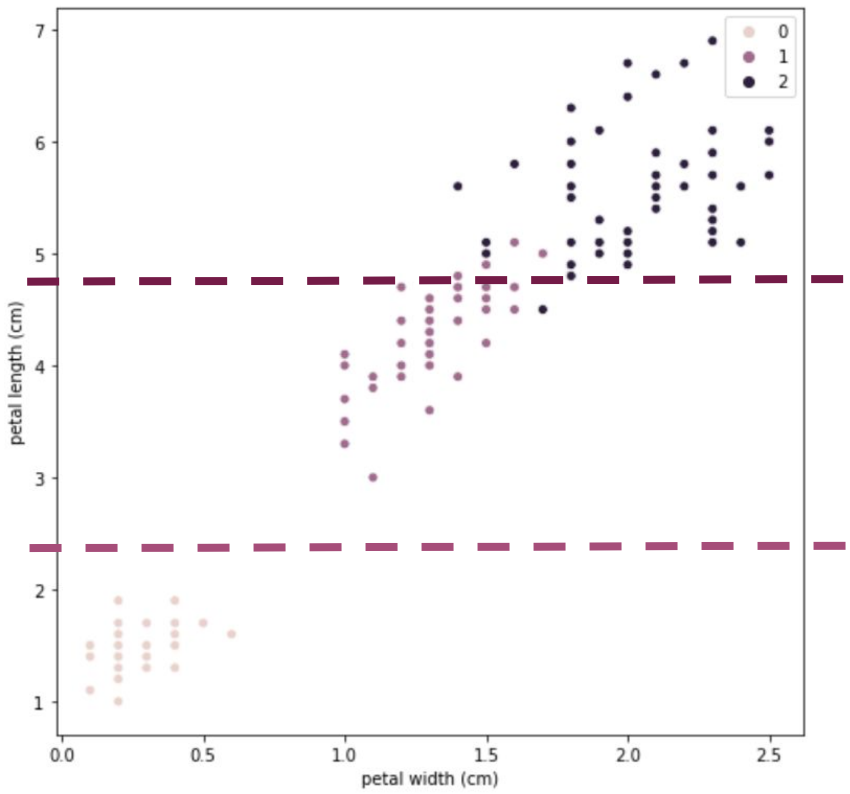
Entropy(PlayGolf) = Entropy (5,9)

= Entropy (0.36, 0.64)

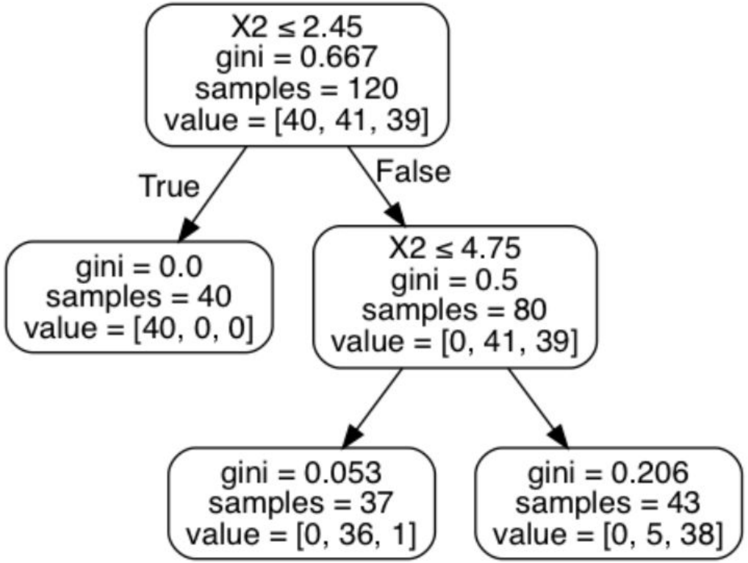
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)

= 0.94

3. building classification trees



- actual results: splits using Gini



4. pros and cons of trees

advantages

- can be used for both classification and regression
- can be displayed graphically/easily interpretable
- non-parametric (does not make any assumptions of the underlying distribution of data), unlike linear regression
- features don't need scaling
- automatically account for interaction

disadvantages

- tend to not perform very well compared to state of the art ML models
- recursive binary splitting makes "locally optimal" decisions that may not result in a globally optimal tree
- easy overfitting