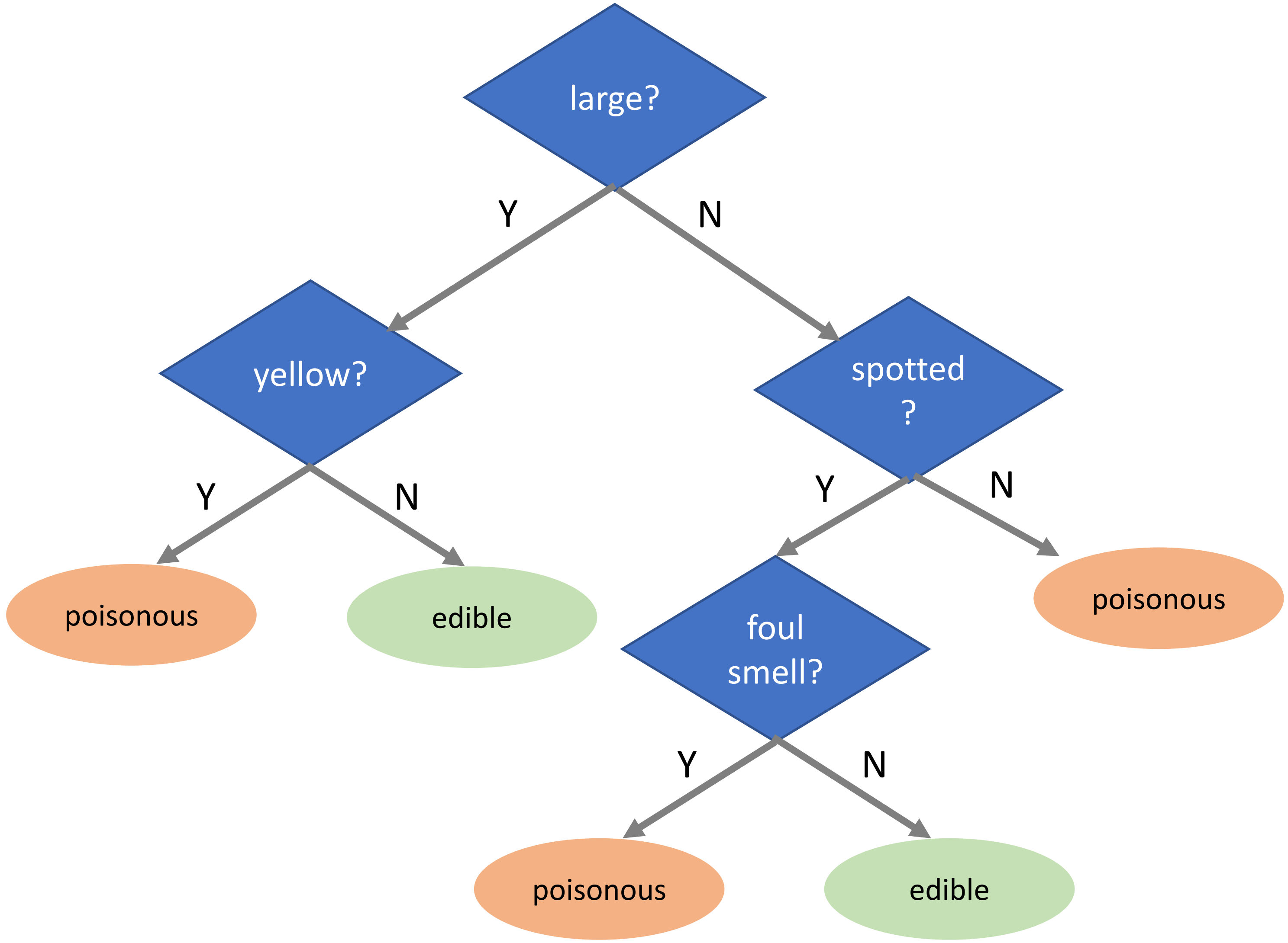
CSCI4622 Machine Learning 2020 Spring

Tree Methods

Geena Kim

*Slide contents adopted from ISLR material

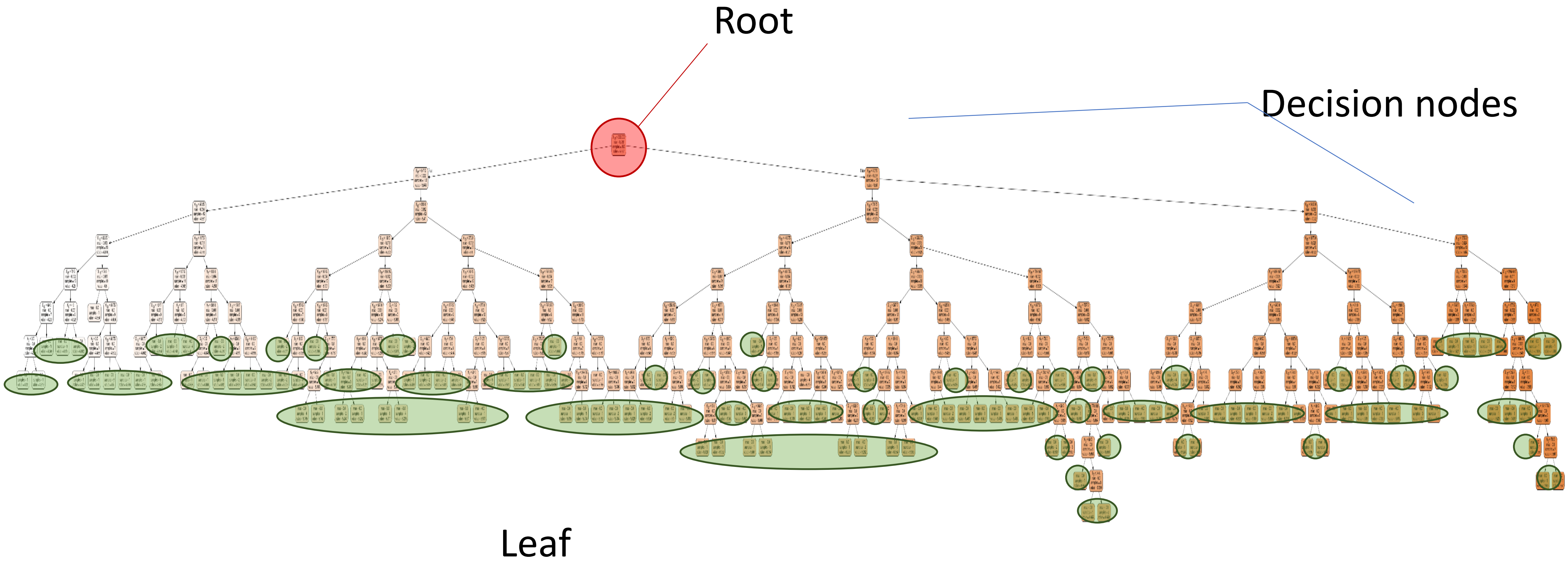
What is Decision Tree?



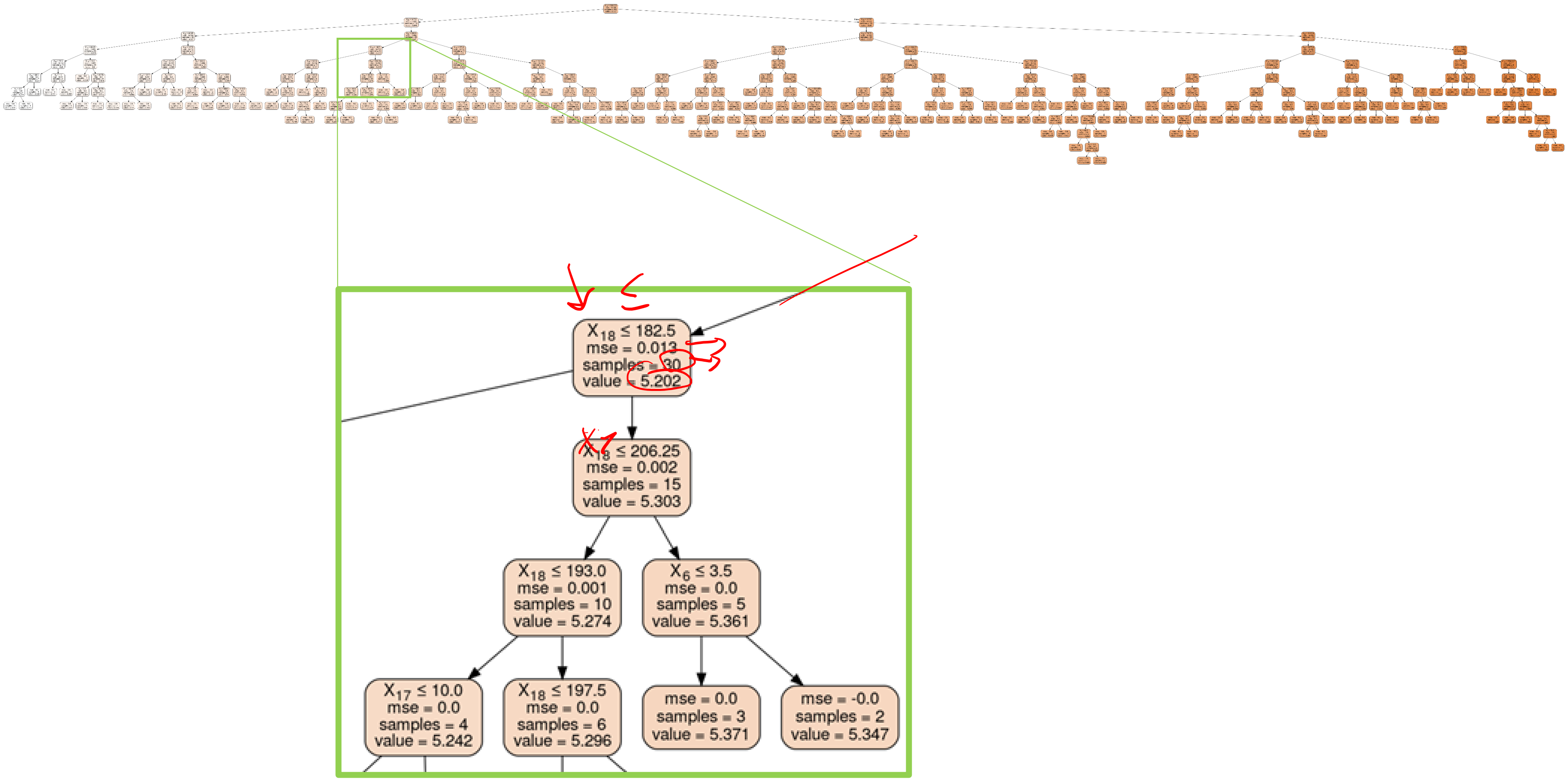
Caesar’s mushroom

Death Cap

Decision Tree Nodes



Decision Nodes



Different kinds of models

Parametric vs. Non-parametric

Linear Regression

Param

$y = ax + b$

Logistic Regression

$z = ax + b$

$\sigma(z) = \frac{1}{1+e^{-z}}$

kNN

Decision Tree

Parameters vs. Hyperparameters

|| Coef

= Design

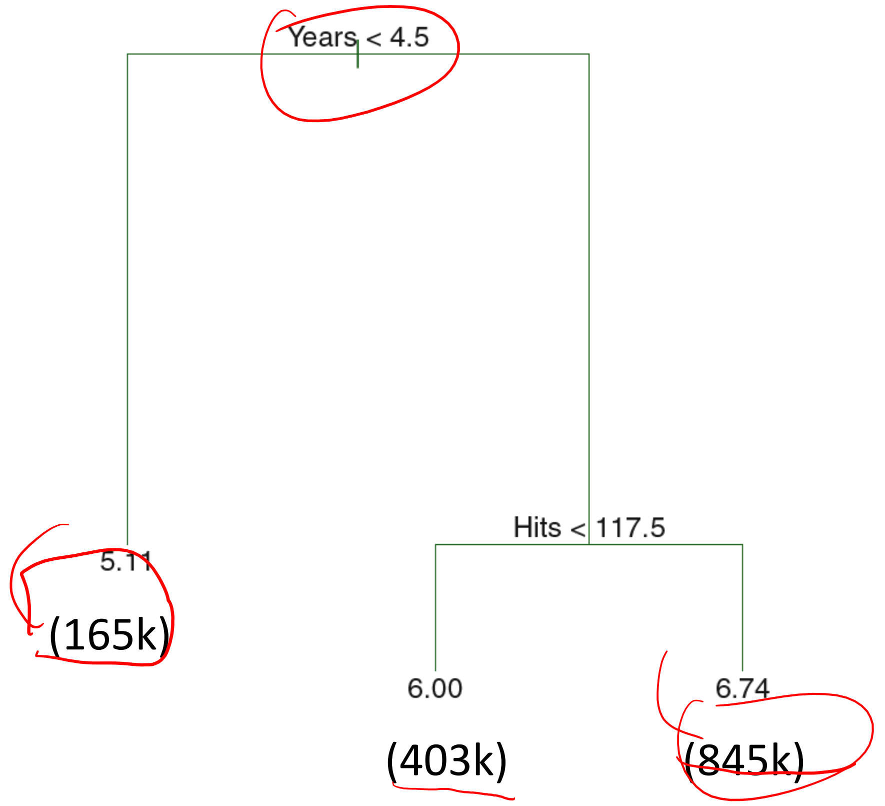
Optimization objective function

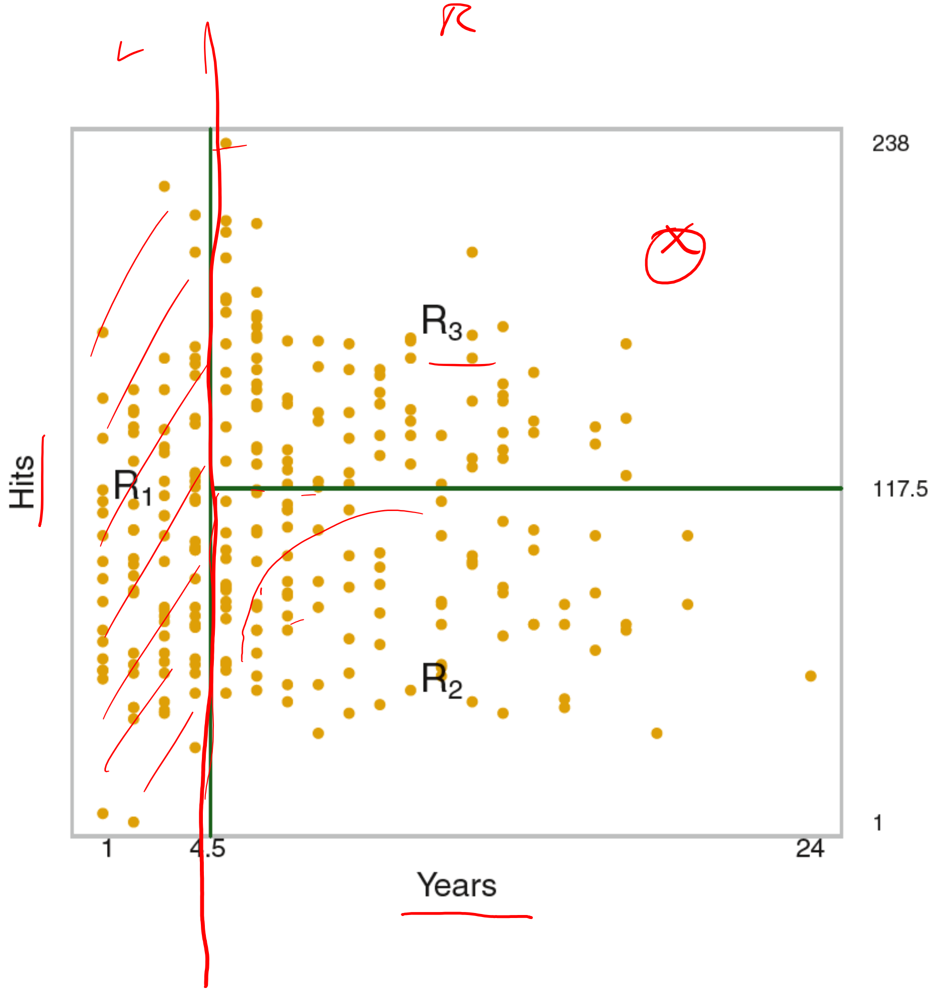
Linear Regression	Minimize MSE
Logistic Regression	Minimize Cross Entropy
kNN	No optimization, but uses distance
Decision Tree	Split to minimize MSE for Regression task and minimize Cross Entropy or Gini for Classification task

Decision Tree Regressor

Predicting Salary of Baseball players

- X1: number of years played in the major league
- X2: number of hits made in the last year
- y: log(salary)



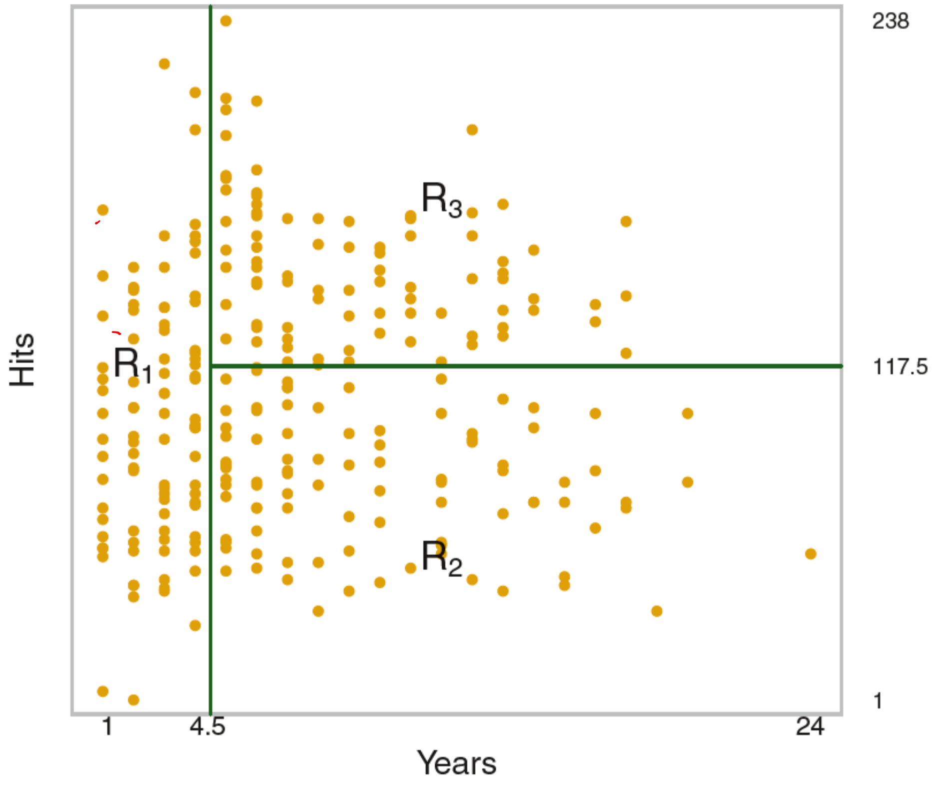


Decision Tree Regressor

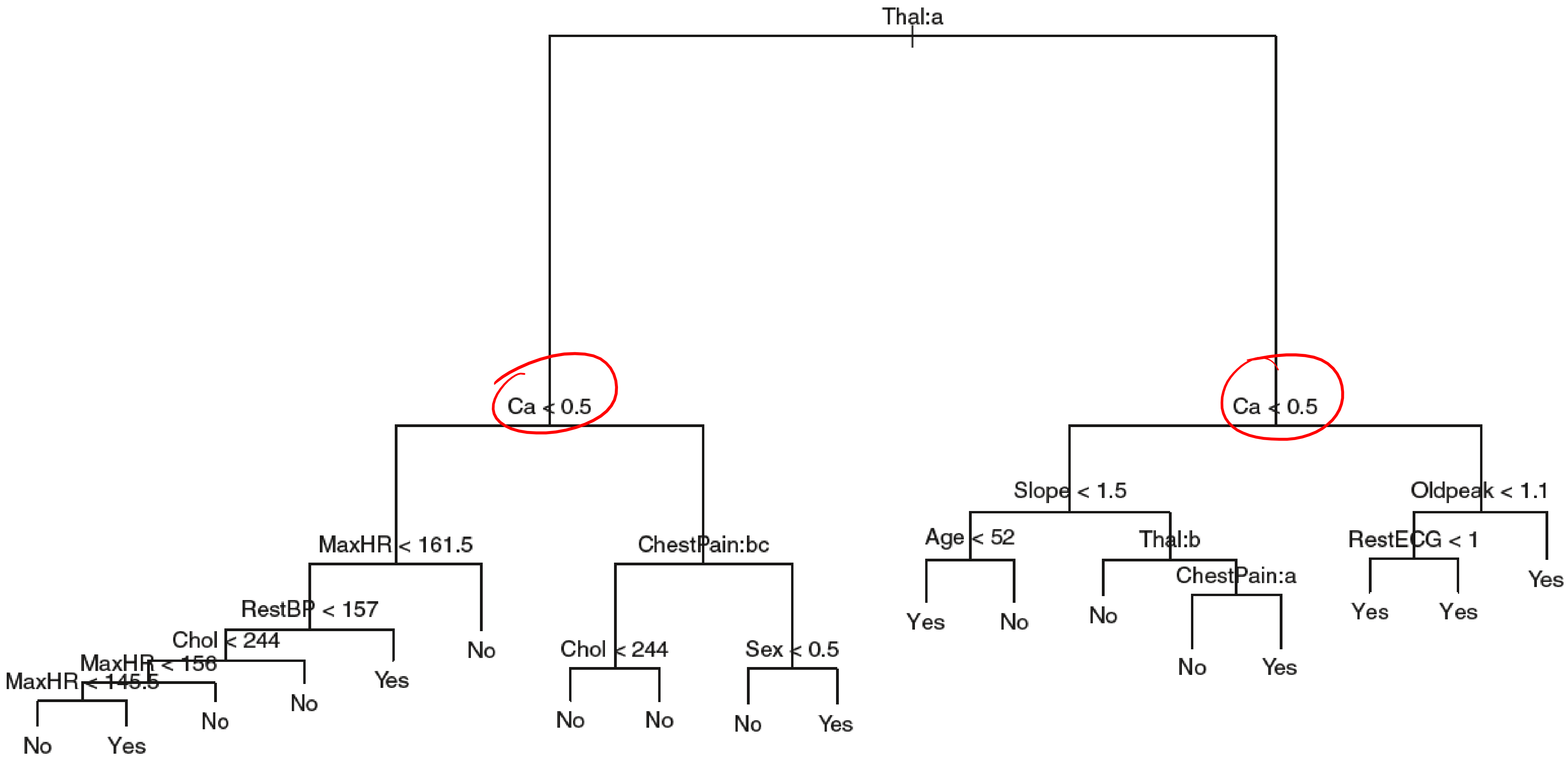
The goal is to find boxes R1 ~ RJ such that

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

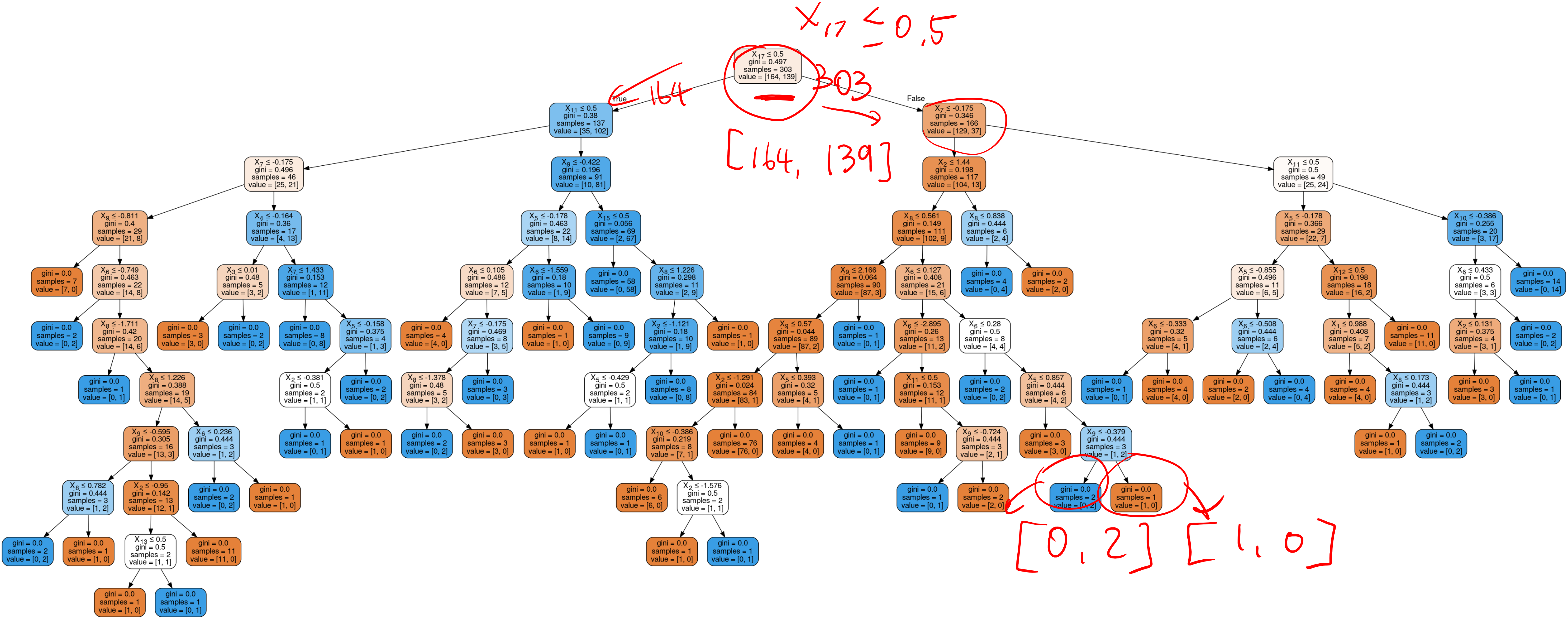
the mean of the data
in the box



Decision Tree Classifier

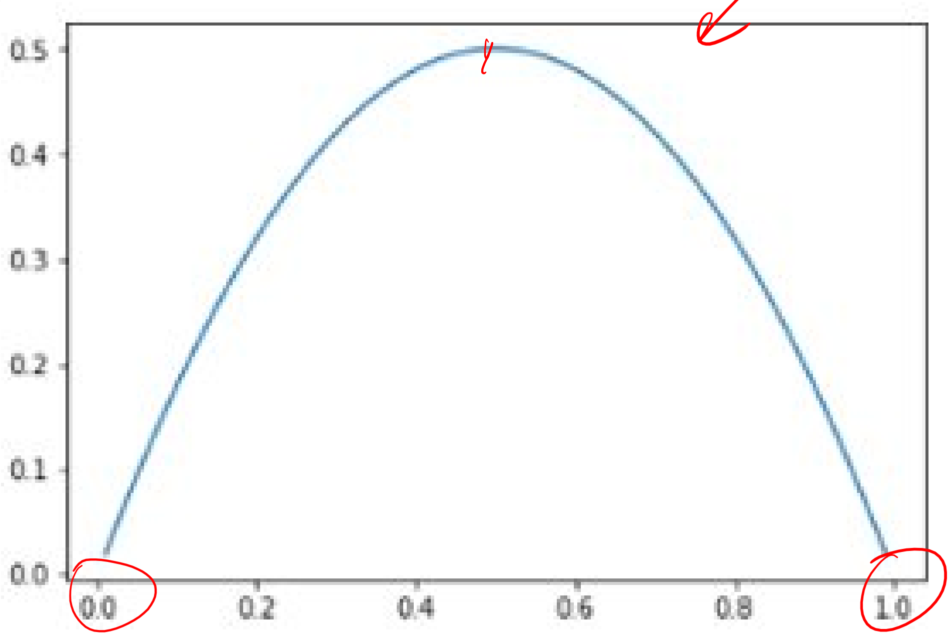


Decision Tree Classifier



Split criterion- Gini index

```
a = np.arange(0.01,1,0.01)
plt.plot(a,2*a*(1-a));
```

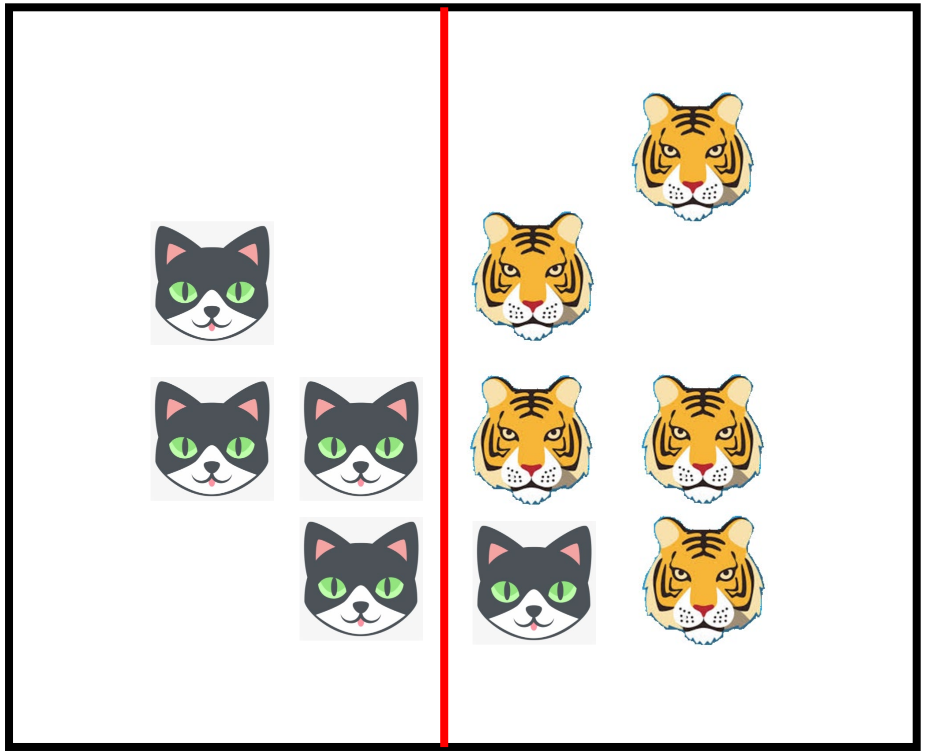


$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

What is the Gini of this box?

Gini: $H(X_m) = \sum_k p_{mk}(1 - p_{mk})$

0,5



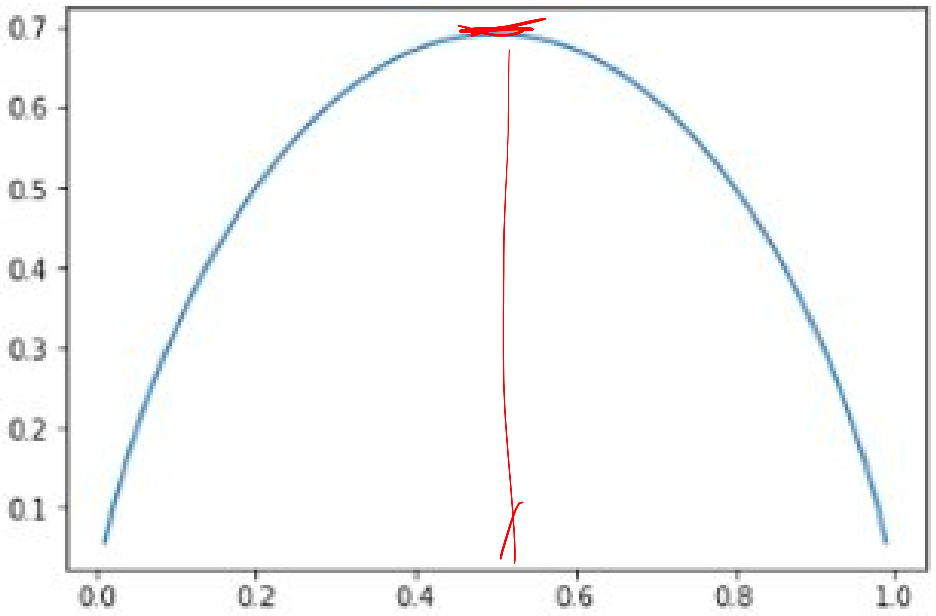
0

$$\underbrace{\frac{1}{6}\cdot\frac{5}{6}}_{\text{cat}} + \underbrace{\frac{5}{6}\cdot\frac{1}{6}}_{\text{tiger}}$$

$$= \frac{5}{18}$$

Split criterion- Entropy

```
a = np.arange(0.01,1,0.01)
plt.plot(a,-a*np.log(a)-(1-a)*np.log(1-a));
```



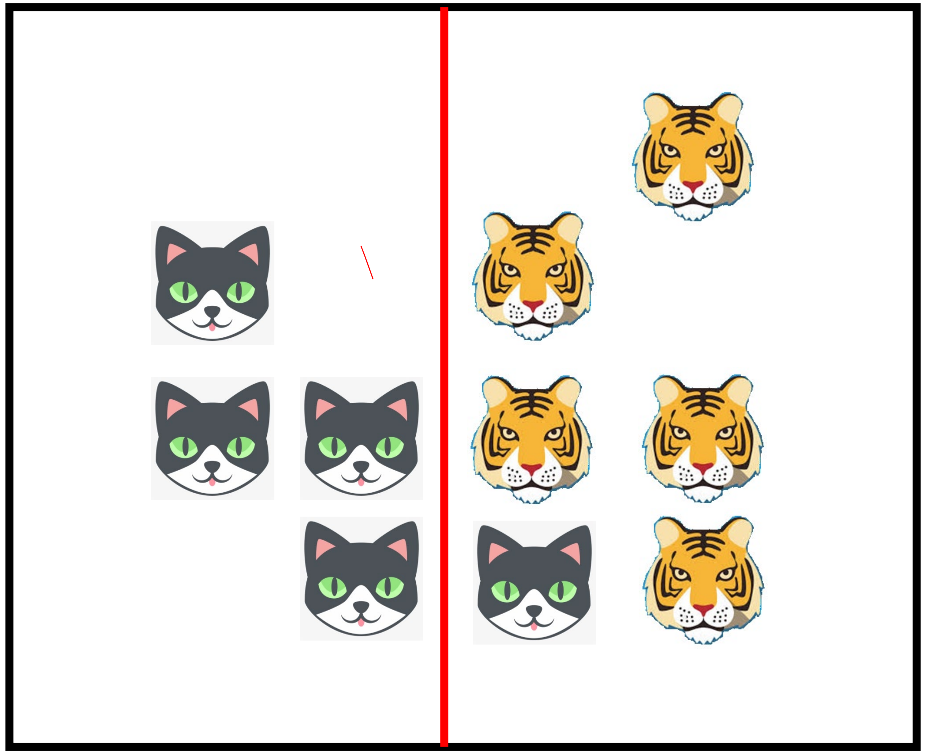
$$H(X_m) = -\sum_k p_{mk} \log(p_{mk})$$

$-\Sigma P \log P$

Split criterion- Information gain

Information Gain = Reduction in Entropy

$$-\left(\frac{1}{2}\log_2\frac{1}{2}+\frac{1}{2}\log_2\frac{1}{2}\right) = 1$$



0

$$-\left(\frac{1}{6}\log_2\frac{1}{6}+\frac{5}{6}\log_2\frac{5}{6}\right) = \underline{0.65}$$

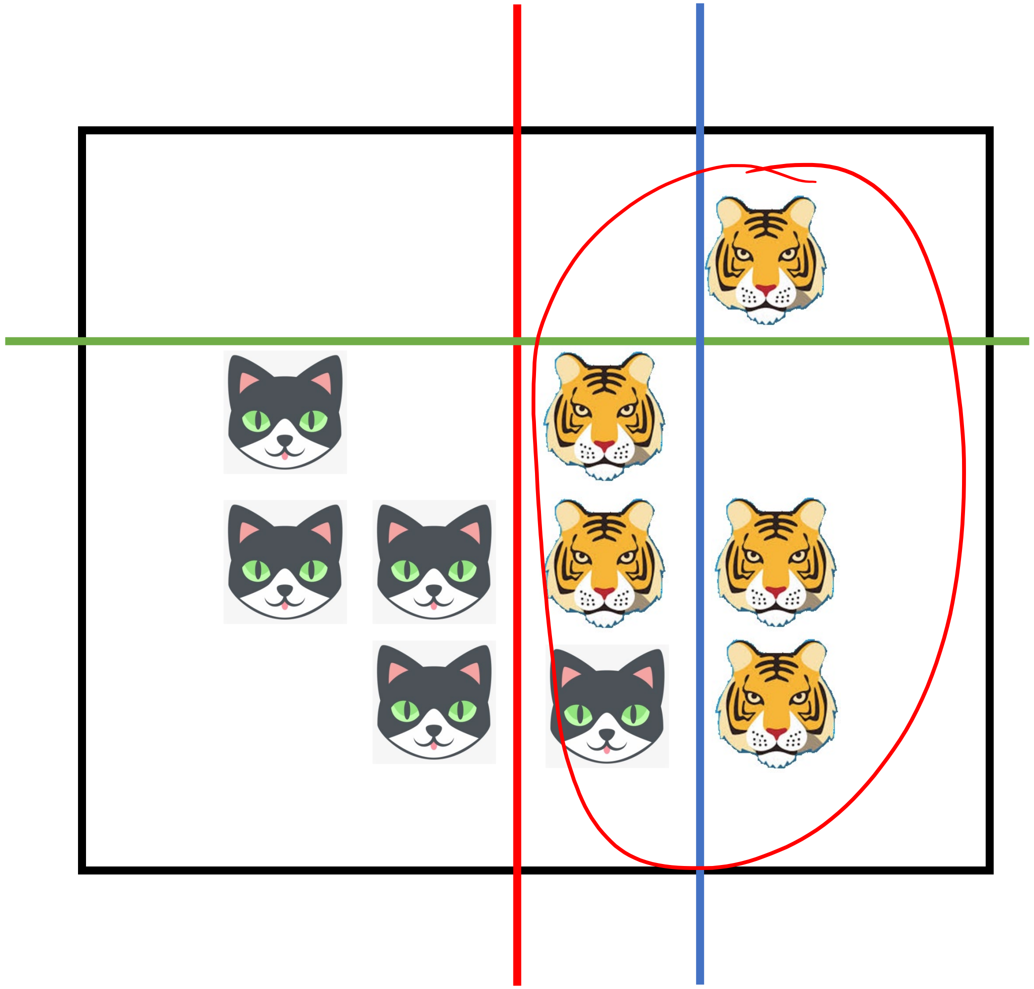
IG

$$= E(\text{parent}) - \frac{N_L}{N_t} \times E_L - \frac{N_R}{N_t} \times E_R$$

$$N_t = N_L + N_R$$

Information Gain = 1 - 0.4*0 - 0.6*0.65 = 0.61

Which split gives the maximum information gain?



Decision Tree Split Criteria

Regression Tree

MSE

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} (y_i - \bar{y}_m)^2$$

MAE

$$H(X_m) = \frac{1}{N_m} \sum_{i \in N_m} |y_i - \bar{y}_m|$$

Classification Tree

Gini

$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

Entropy

$$H(X_m) = -\sum_k p_{mk} \log(p_{mk})$$

Information Gain = E(parent)-E(children)

Decision Tree – When to stop split?

max_depth The maximum depth of the tree

min_samples_split The minimum number of samples required to split an internal node

min_samples_leaf The minimum number of samples required to be at a leaf node

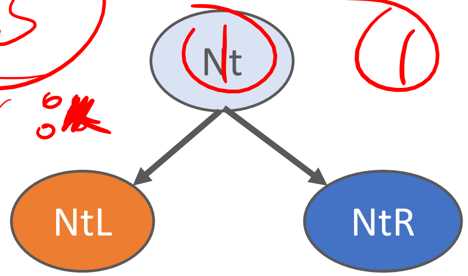
max_features The number of features to consider when looking for the best split

→ sqrt

min_impurity_decrease A node will be split if this split induces a decrease of the impurity greater than or equal to this value

The weighted impurity decrease equation is the following:

```
N_t / N * (impurity - N_t_R / N_t * right_impurity
                    - N_t_L / N_t * left_impurity)
```



Hyperparameter search

Grid Search Tip

- Give a range of values for each hyperparameter
- Measure a training time for one, then estimate how long for the loop
- Adjust number of values, range, or hyperparameters to include

[5, 10, 15, 20]

5, 6, 7, 8, 9, - - -

max_depth = [5, 10, 15, 20]

min_samples_split = [5, 6, 7]

min_samples_leaf

max_features

min_impurity_decrease

Decision Tree Pros and Cons

Trees are easy to understand

Trees don't suffer collinearity

Trees are good for non-linear features

Trees handle categorical variables easily

Trees are weak-learner

Trees have high variance in general

Linear regression is a better choice if features are linear

Tree's performance can be greatly improved when ensembled