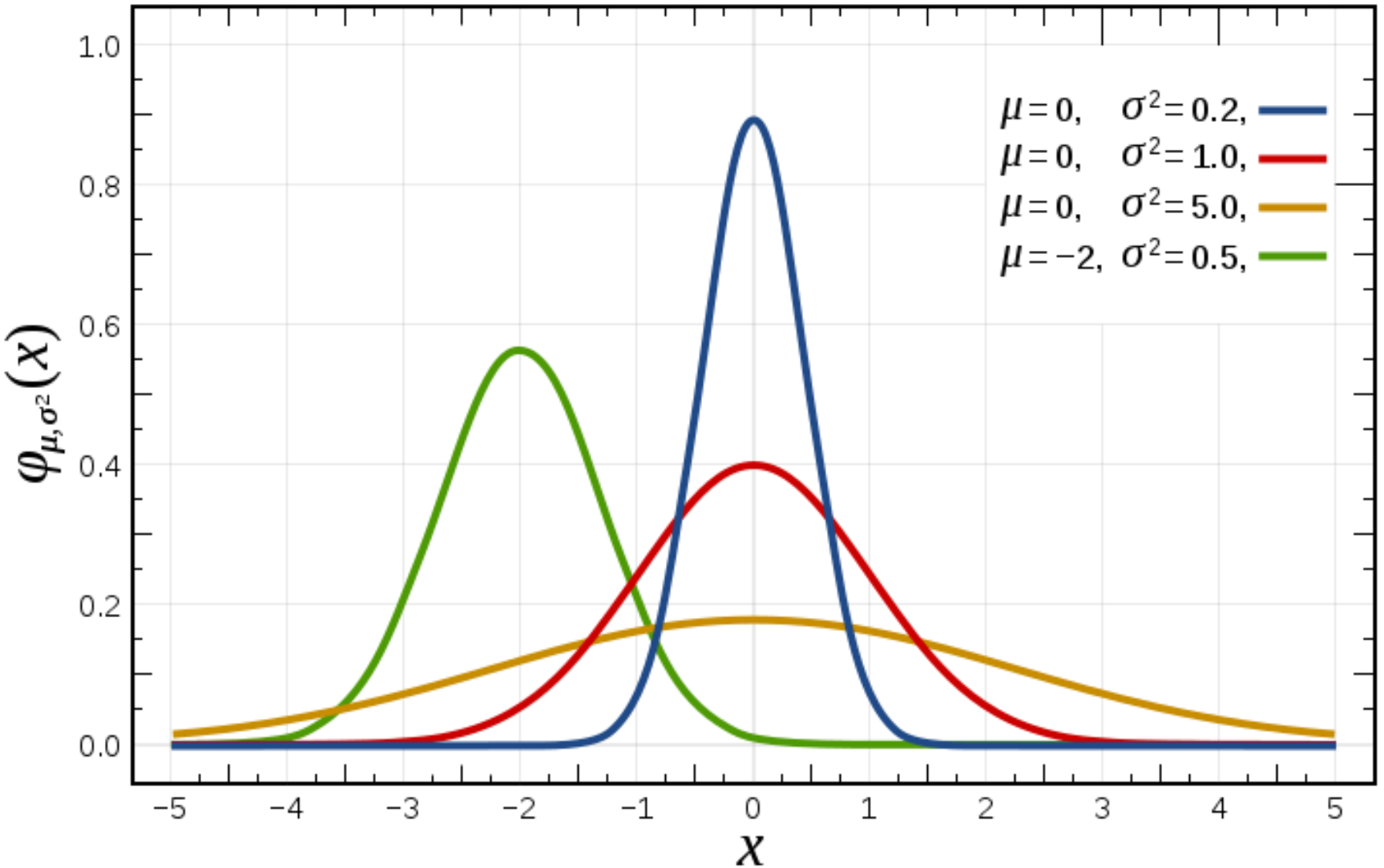
Machine Learning / Tree based methods

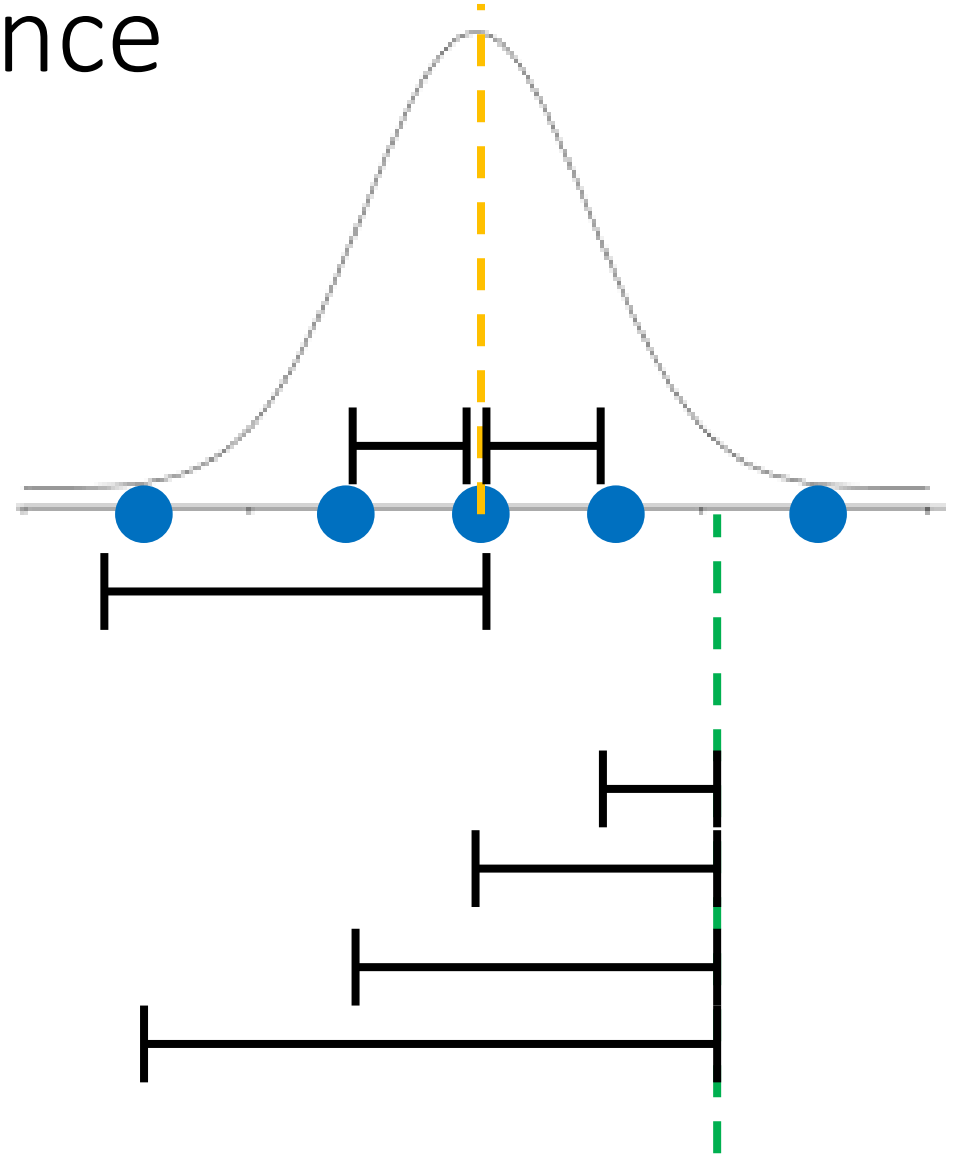
Decision Trees

Regression

Variance


1.0
0.8
0.6
0.4
0.2
0.0
$\varphi_{\mu,\sigma^2}(x)$
−5
−4
−3
−2
−1
0
1
2
3
4
5
x
$\mu = 0, \quad \sigma^2 = 0.2,$
$\mu = 0, \quad \sigma^2 = 1.0,$
$\mu = 0, \quad \sigma^2 = 5.0,$
$\mu = -2, \quad \sigma^2 = 0.5,$

Variance

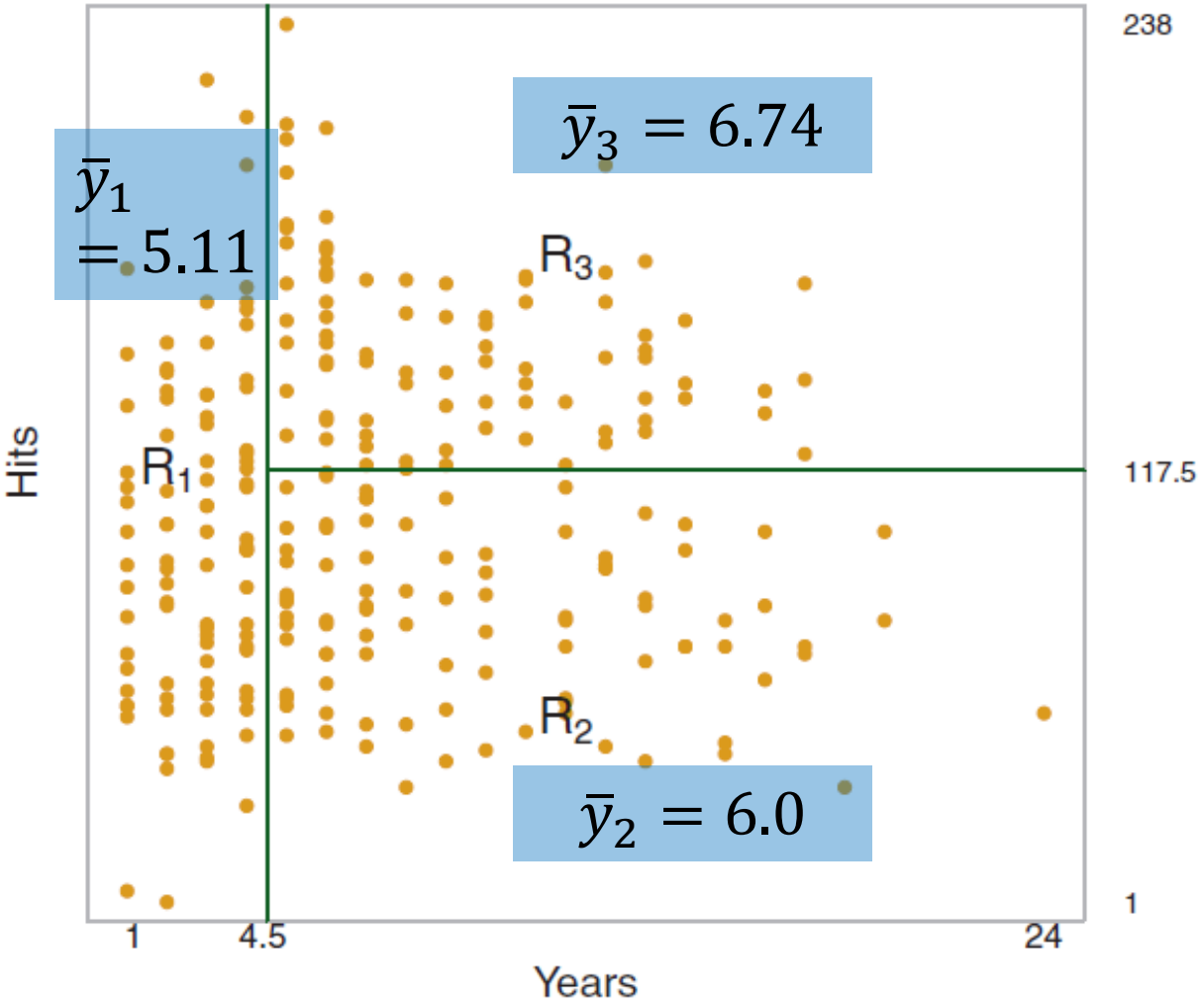


$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - c)^2$$

$c = \bar{x}$ produce the lowest value of σ^2

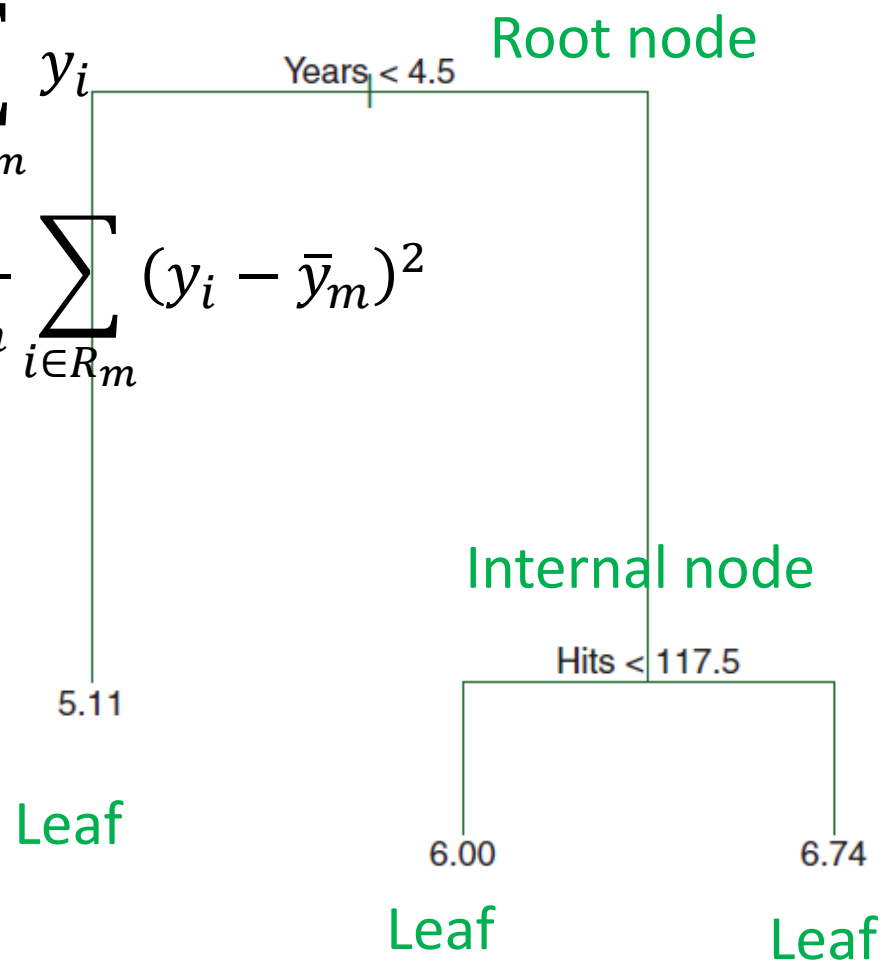
Domain partitioning

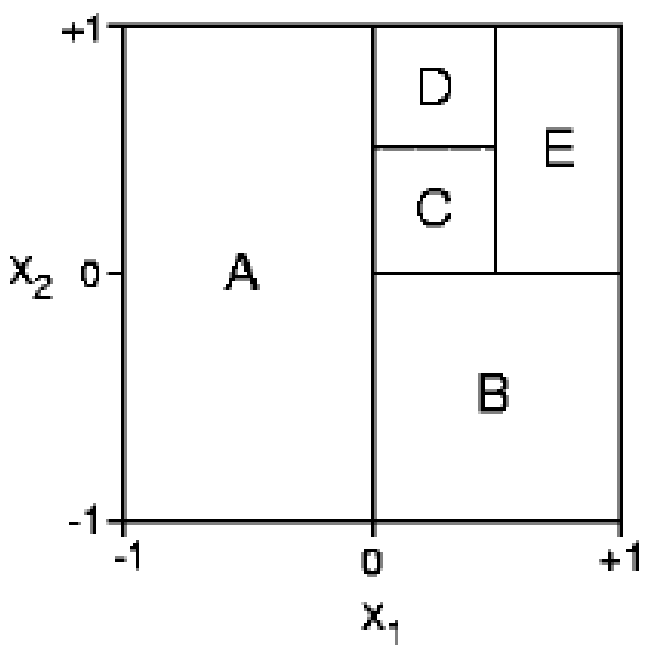
$$f(x_1, x_2) = y$$

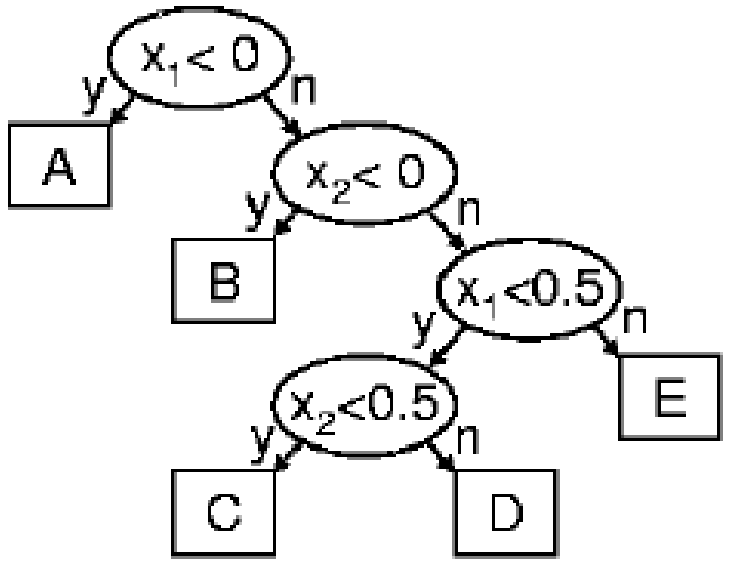


$$\bar{y}_m = \frac{1}{N_m} \sum_{i \in R_m} y_i$$

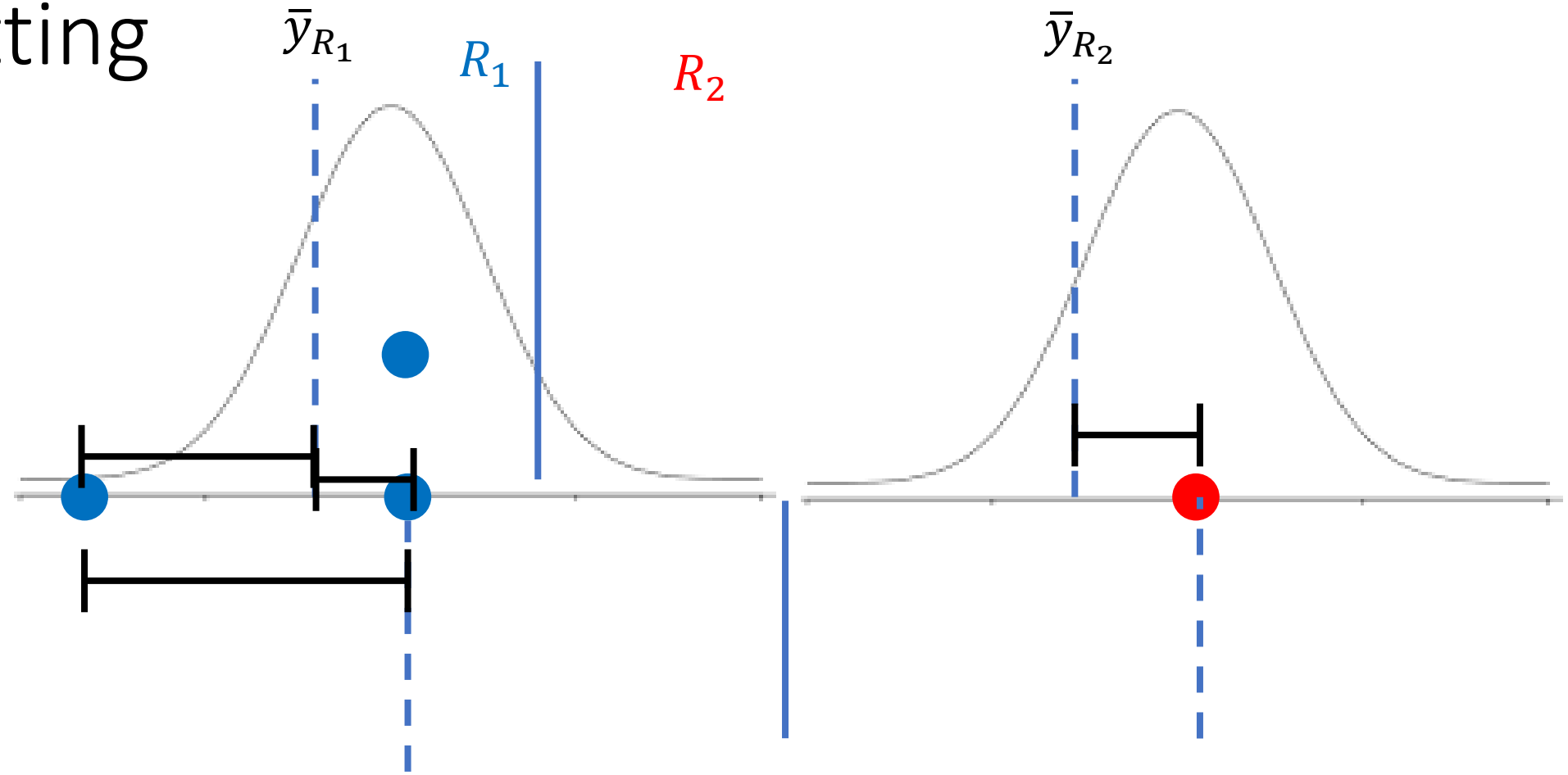
$$H(X_m) = \frac{1}{N_m} \sum_{i \in R_m} (y_i - \bar{y}_m)^2$$




+1
D
E
C
A
x_2 0
B
-1
-1
0
+1
x_1


$x_1 < 0$
y
n
A
$x_2 < 0$
y
n
B
$x_1 < 0.5$
y
n
E
$x_2 < 0.5$
y
n
C
D

Fitting

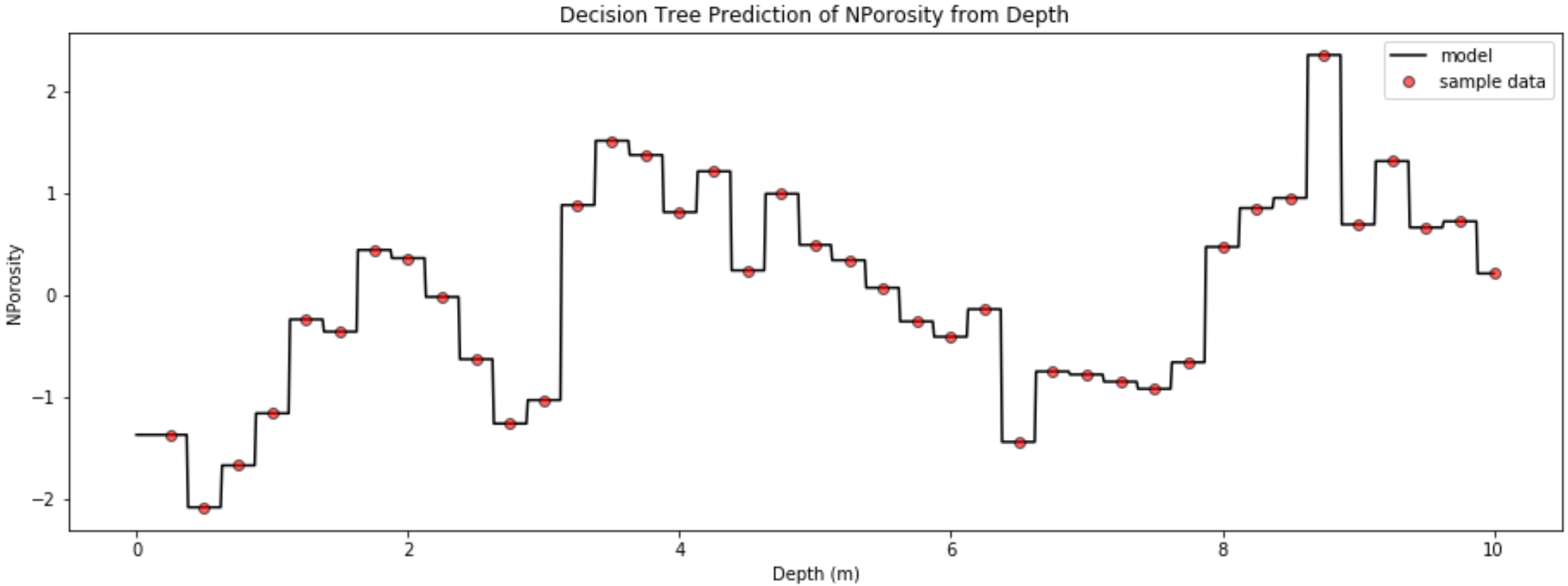


Fit

Step 1,find j and s that Minimize:

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

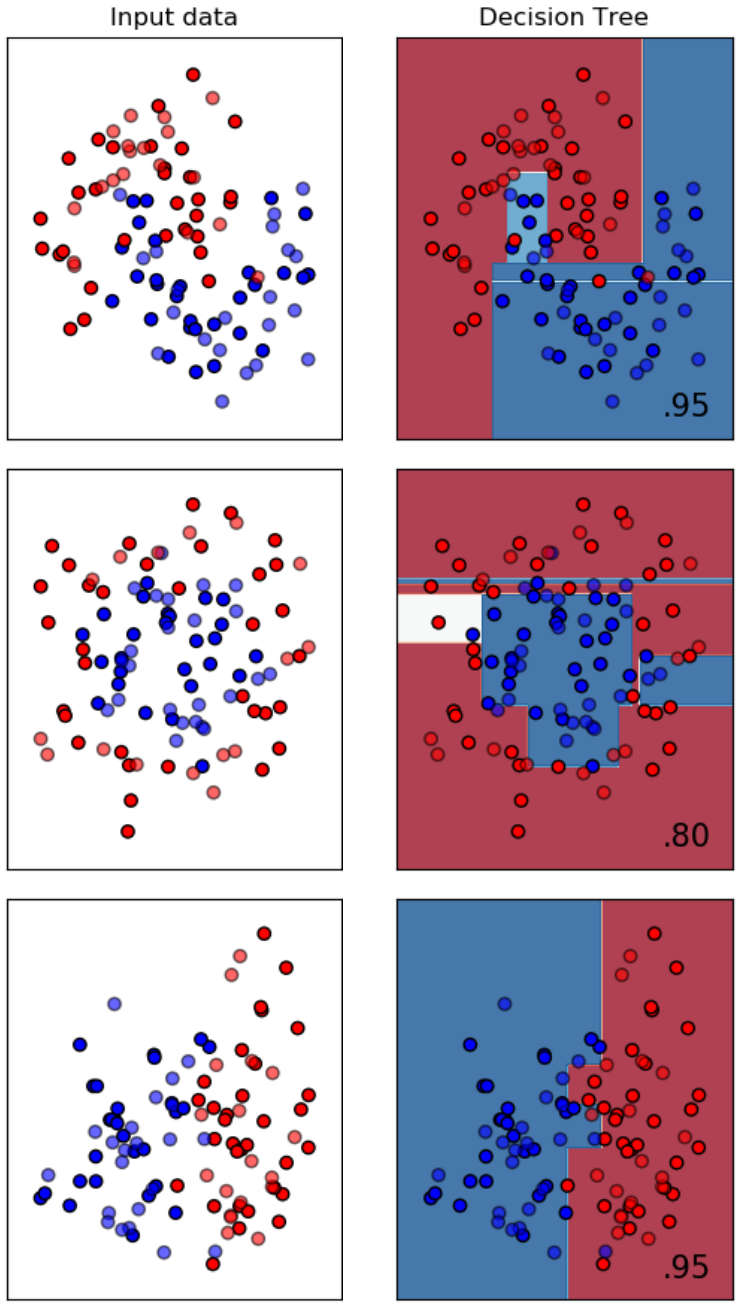
$j = 1, \ldots, n$
$s =$ cuttoff



https://github.com/GeostatsGuy/PythonNumericalDemos/blob/master/SubsurfaceDataAnalytics_PolygonalRegression.ipynb

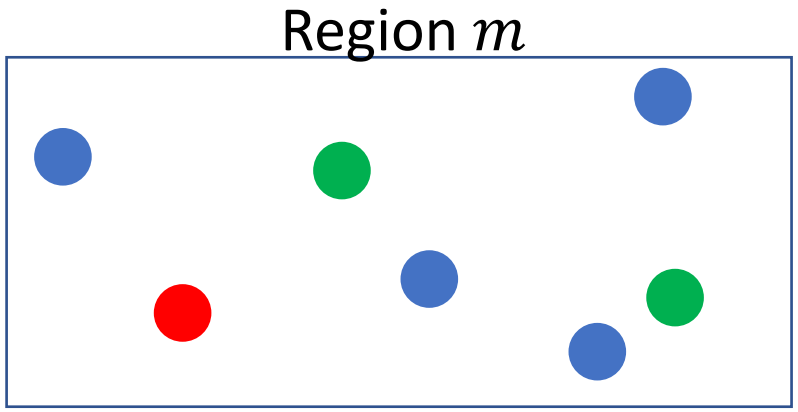
Decision Trees

Classification



Proportions

$$p_{mk} := \frac{1}{N_m} \sum_{x_i \in R_m} \mathbb{I}(y_i = k)$$



$$N = 7$$

$$p_{m1} = \frac{1}{7}$$

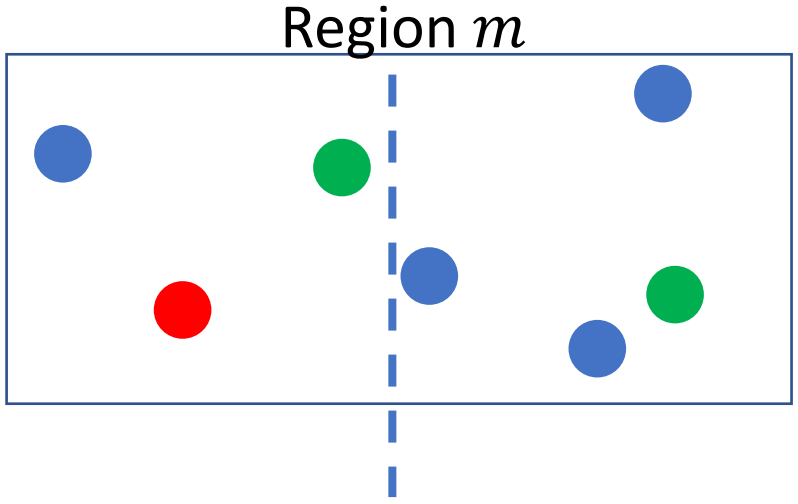
$$p_{m2} = \frac{2}{7}$$

$$p_{m3} = \frac{4}{7}$$

https://scikit-learn.org/stable/modules/tree.html#classification-criteria

p. 311-313, Hastie, 2014

Impurity functions. Misclassification



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

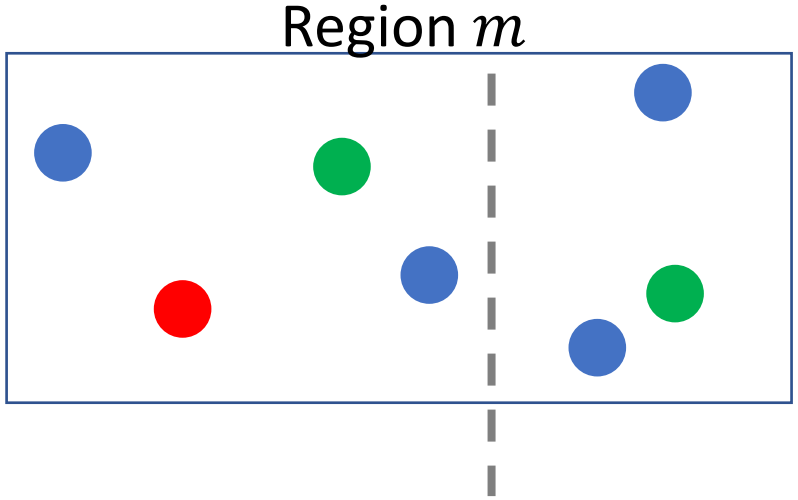
Misclassification

$$\frac{2}{3} + \frac{1}{4} = \frac{11}{12}$$

https://scikit-learn.org/stable/modules/tree.html#classification-criteria

p. 311-313, Hastie, 2014

Impurity functions. Misclassification



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

Misclassification

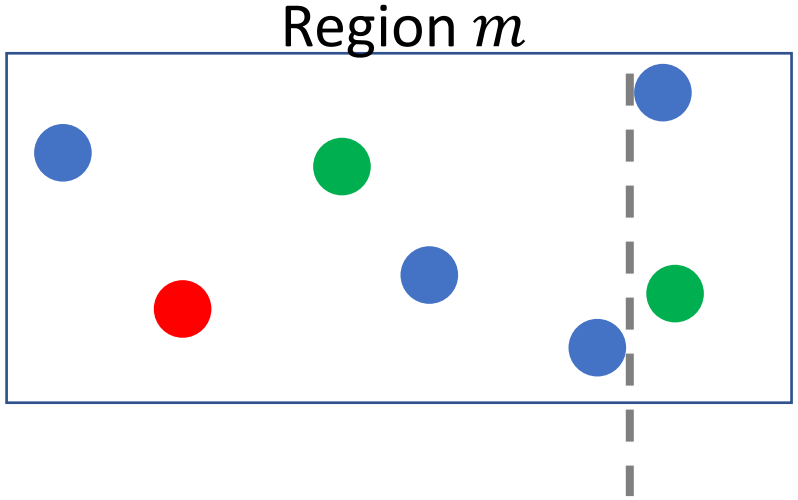
$$\frac{2}{3} + \frac{1}{4} = \frac{11}{12}$$

$$\frac{1}{2} + \frac{1}{3} = \frac{5}{6}$$

https://scikit-learn.org/stable/modules/tree.html#classification-criteria

p. 311-313, Hastie, 2014

Impurity functions. Misclassification



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

Misclassification

$$\frac{2}{3} + \frac{1}{4} = \frac{11}{12}$$

$$\frac{1}{2} + \frac{1}{3} = \frac{5}{6}$$

https://scikit-learn.org/stable/modules/tree.html#classification-criteria

p. 311-313, Hastie, 2014

Impurity functions. Gini Index

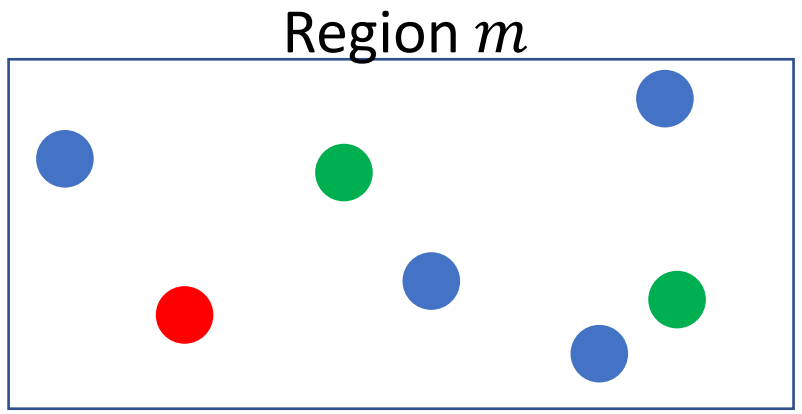
$$H(X_m) = \sum_k p_{mk}(1 - p_{mk})$$

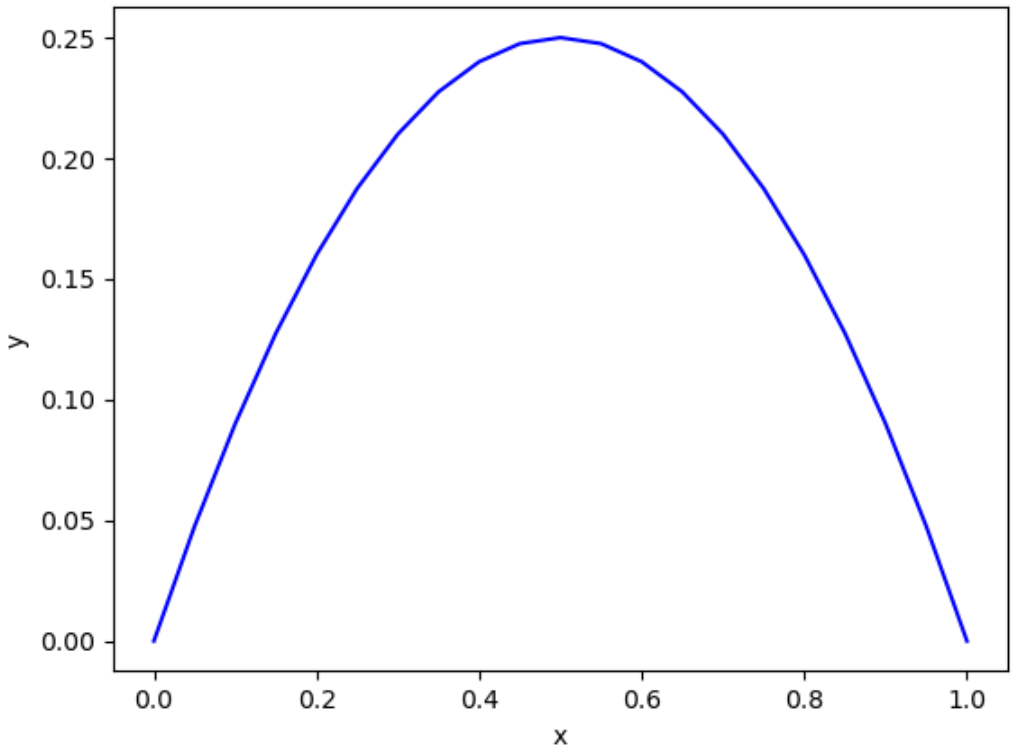
- Find the maximum of $y = x(1 - x)$
- Plot it

If only 2 classes, consider the following 2 cases

$a) p_1 \gg p_2$

$b) p_1 \approx p_2$

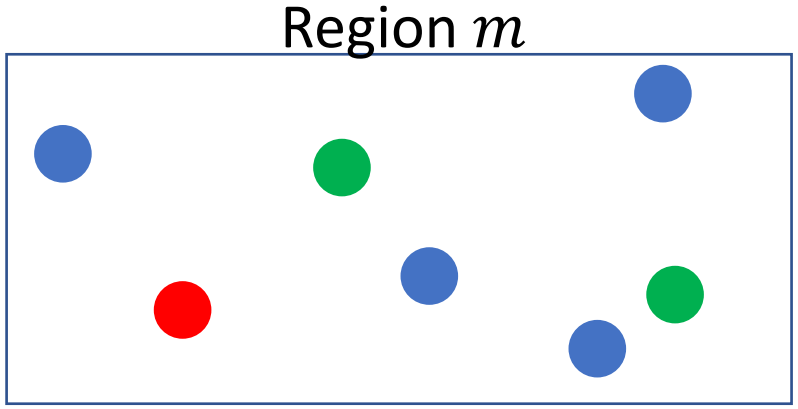




https://scikit-learn.org/stable/modules/tree.html#classification-criteria

p. 311-313, Hastie, 2014

Impurity functions



$$p_{mk} := \frac{1}{N} \sum_{x_i \in R_m} \mathbb{I}(y_i = k)$$

$$H(X_m) = 1 - \max_k(p_{mk}) \qquad \text{Misclassification}$$

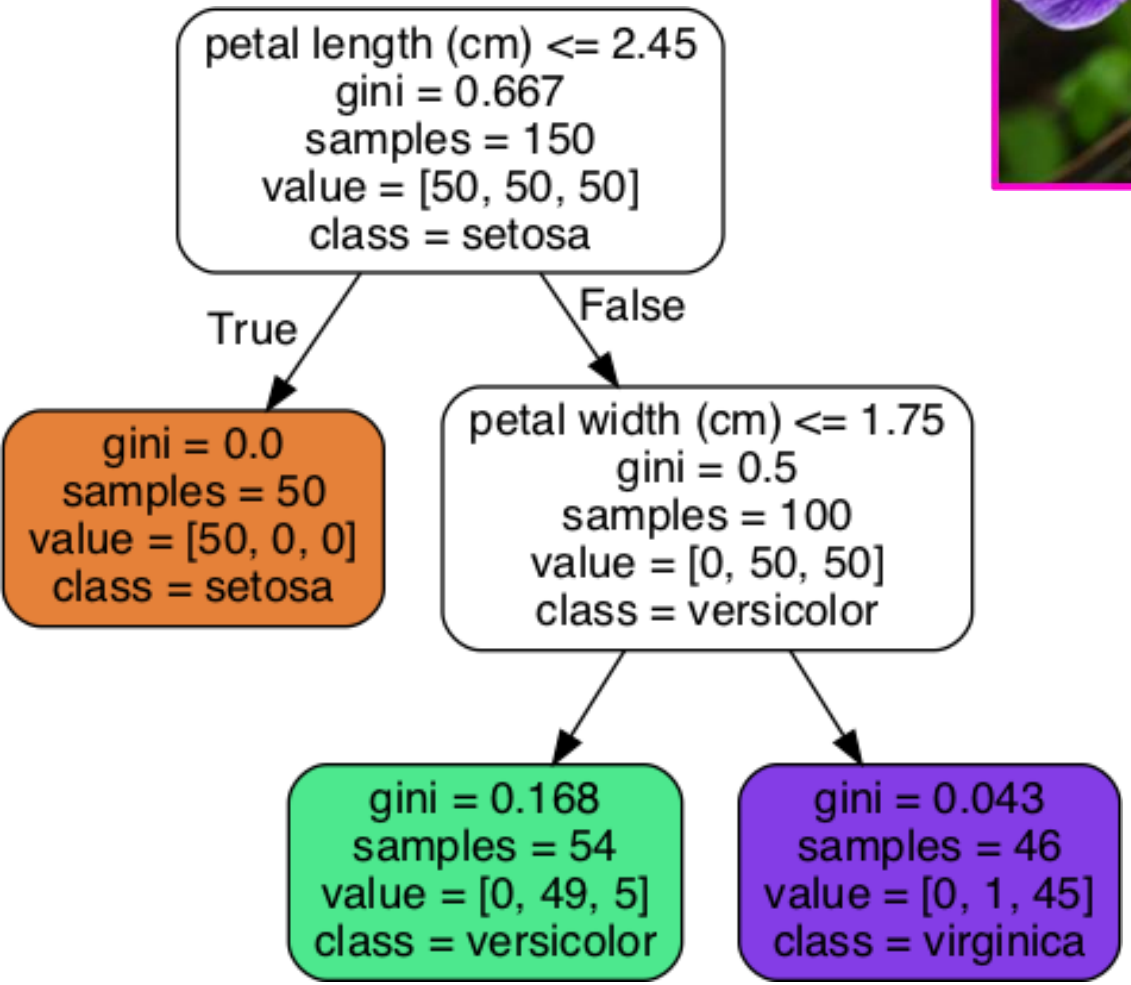
$$H(X_m) = \sum_k p_{mk}(1 - p_{mk}) \qquad \text{Gini Index}$$

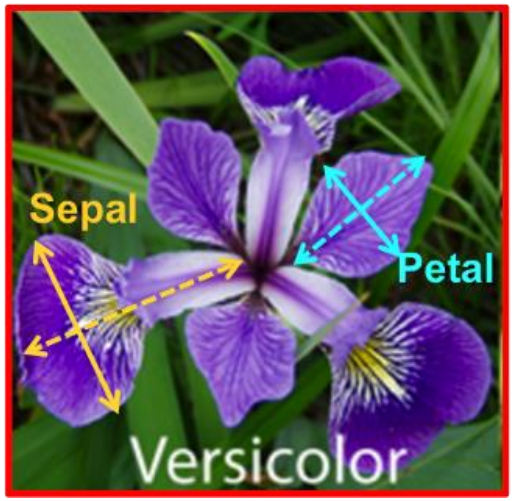
$$H(X_m) = -\sum_k p_{mk} \log(p_{mk}) \qquad \text{Entropy}$$

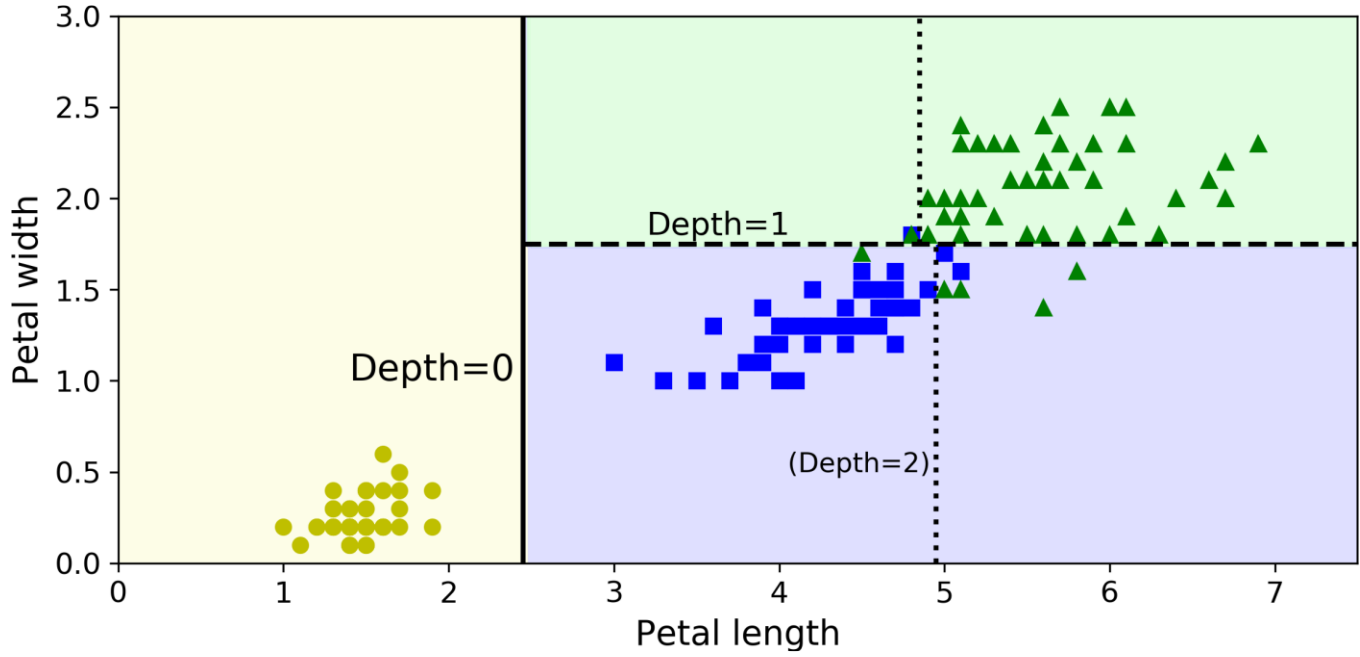
https://scikit-learn.org/stable/modules/tree.html#classification-criteria

p. 311-313, Hastie, 2014

Iris dataset







Homework assigment

Train and fine-tune a Decision Tree for the moons dataset.

a. Generate a moons dataset using `make_moons(n_samples=10000, noise=0.4)`.

b. Split it into a training set and a test set using `train_test_split()`.

c. Use grid search with cross-validation (with the help of the `GridSearchCV` class) to find good hyperparameter values for a `DecisionTreeClassifier`.

Hint: try various values for `max_leaf_nodes`.

d. Train it on the full training set using these hyperparameters, and measure your model's performance on the test set. You should get roughly 85% to 87% accuracy.