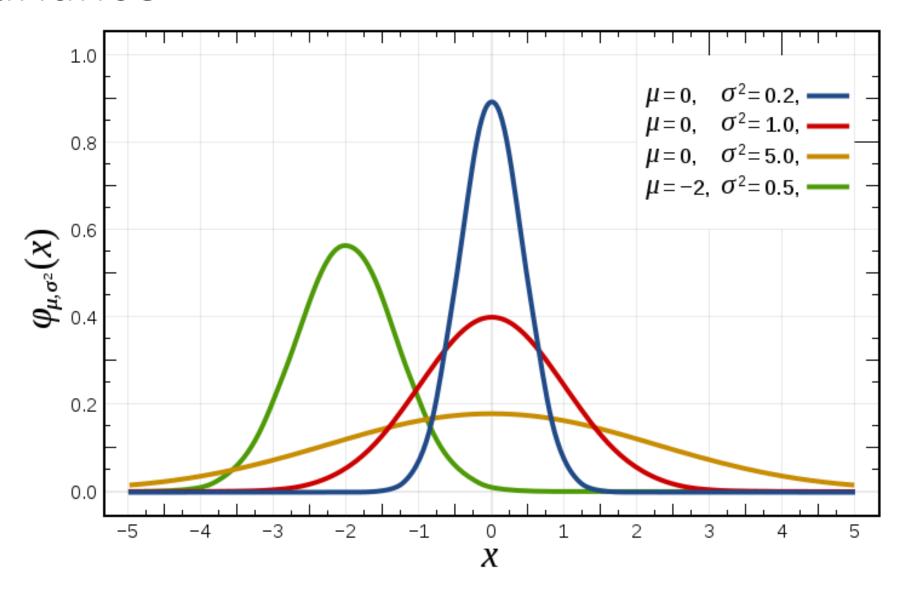
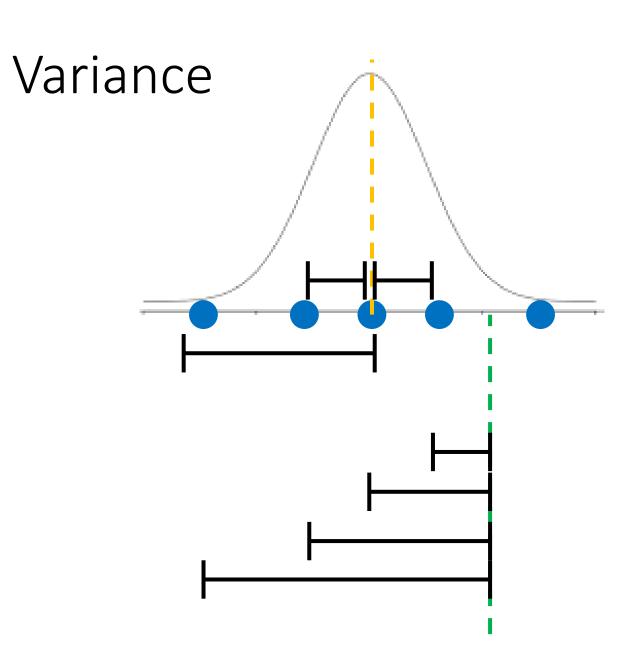
Machine Learning / Tree based methods

Decision Trees

Regression

Variance



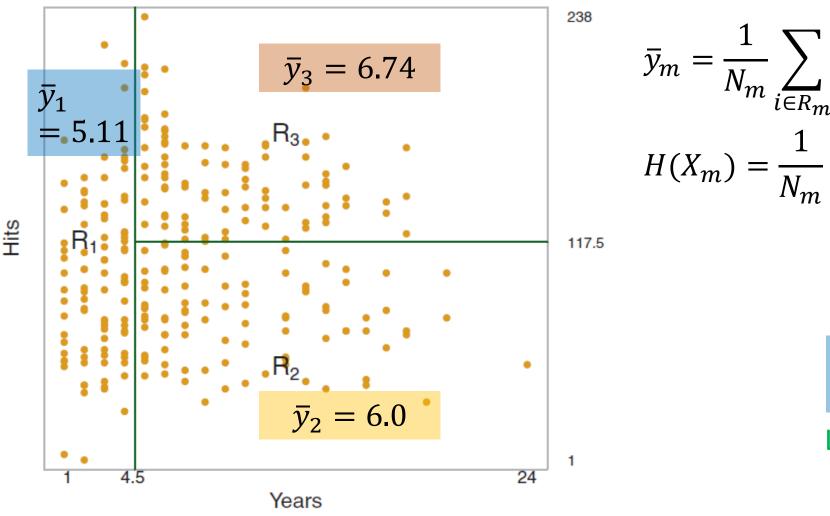


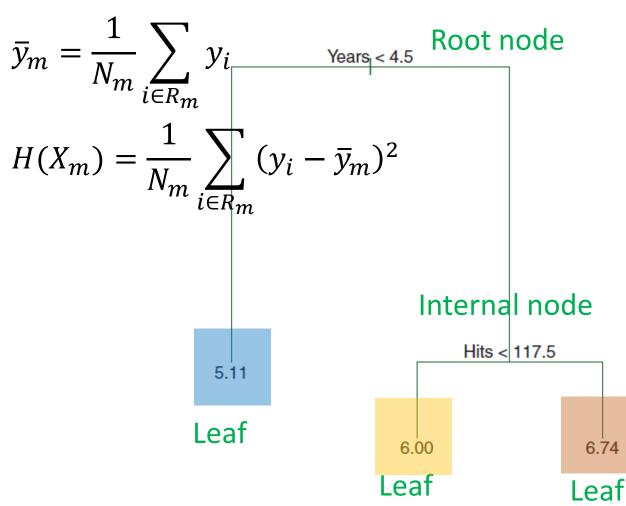
$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - c)^2$$

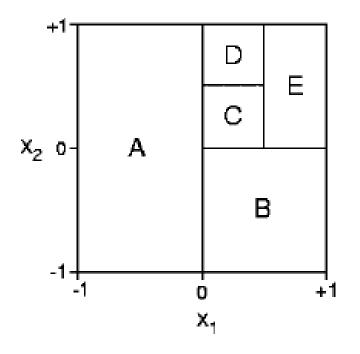
 $c=\bar{x}$ produce the lowest value of σ^2

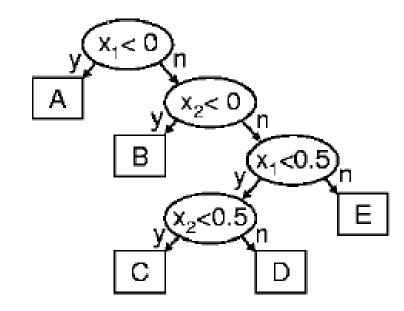
Domain partitioning

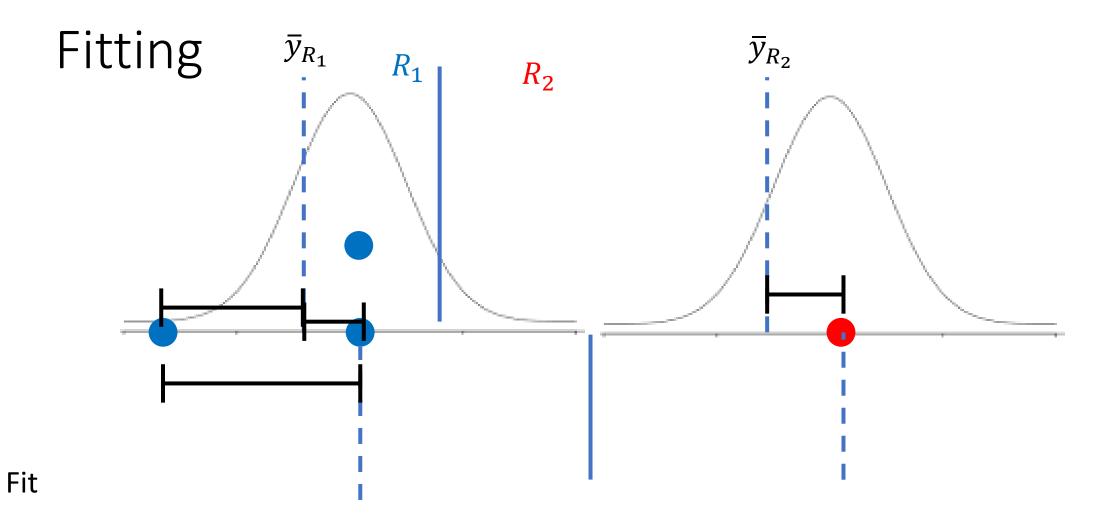
$$f(x_1, x_2) = y$$







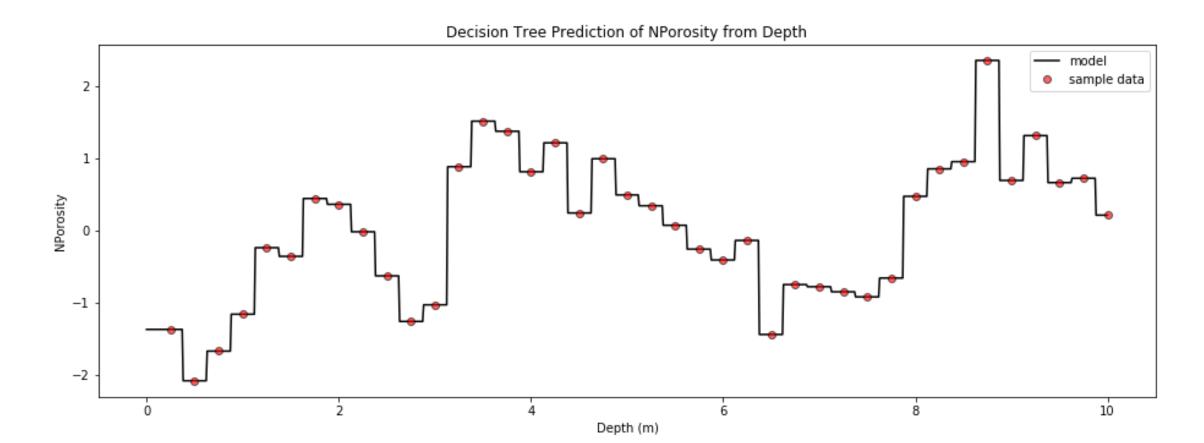




Step 1, find *j* and *s* that Minimize:

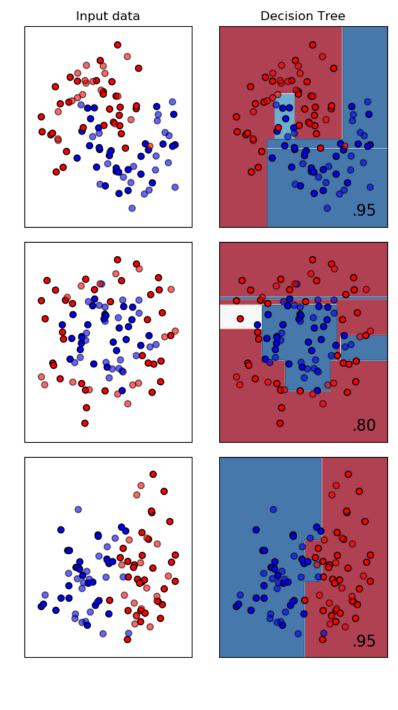
$$\sum_{i:x_i \in R_1(j,s)} (y_i - \bar{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \bar{y}_{R_2})^2$$

j = 1, ..., ns = cuttoff



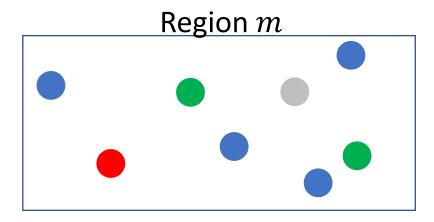
Decision Trees

Classification



Proportions

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



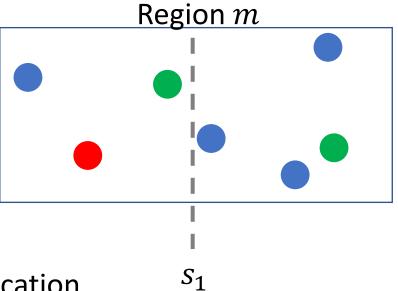
$$N = 7$$

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

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

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

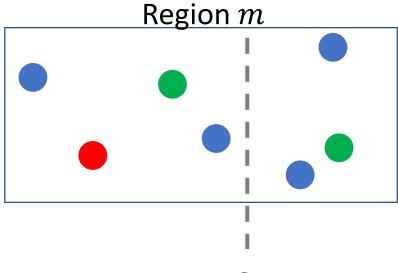
Misclassification measure



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

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

Misclassification measure



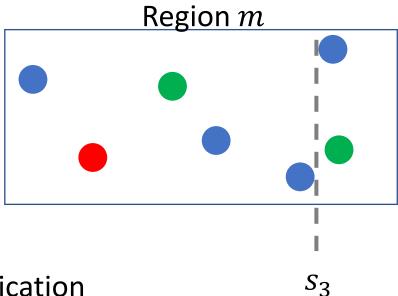
$$H(X_m) = 1 - \max_k(p_{mk})$$

$$S_2$$

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

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

Misclassification measure



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

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

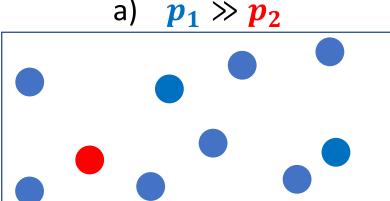
$$\frac{1}{2} + \frac{1}{3} = \frac{5}{6}$$

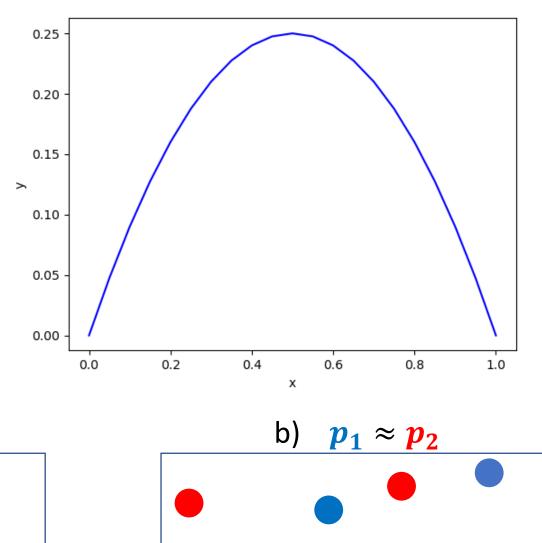
Impurity functions. Gini Index

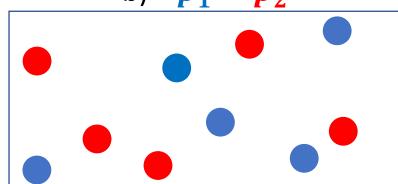
$$H(X_m) = \sum_k p_{mk} (1 - p_{mk})$$

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

If only 2 classes, consider the following 2 cases







Impurity functions

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

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

Misclassification

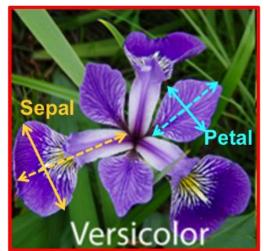
$$H(X_m) = \sum_{k} p_{mk} (1 - p_{mk})$$
 Gini Index

$$H(X_m) = -\sum_{l} p_{mk} \log(p_{mk})$$
 Entropy

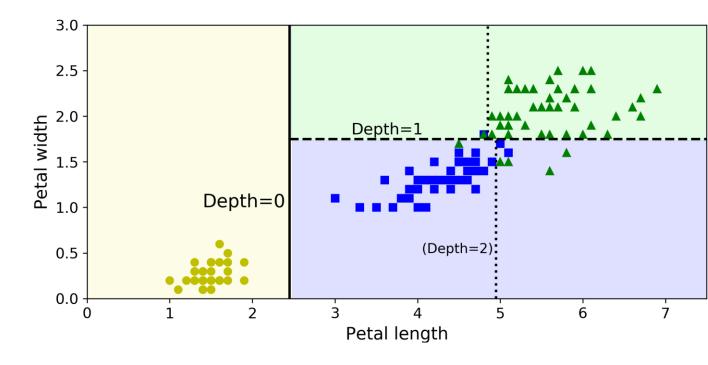
Iris dataset

```
petal length (cm) <= 2.45
               gini = 0.667
              samples = 150
           value = [50, 50, 50]
              class = setosa
                            False
         True
                      petal width (cm) <= 1.75
   gini = 0.0
                              gini = 0.5
 samples = 50
                           samples = 100
value = [50, 0, 0]
                         value = [0, 50, 50]
 class = setosa
                          class = versicolor
                 gini = 0.168
                                        gini = 0.043
                samples = 54
                                       samples = 46
               value = [0, 49, 5]
                                     value = [0, 1, 45]
              class = versicolor
                                      class = virginica
```









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.