CLASSIFICATION II
DECISION TREE

Dr. Lablanche Pierre-Yves
African Institute for Mathematical Sciences

INTRODUCTION

Highly recommend the following link for a more visual introduction :

http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

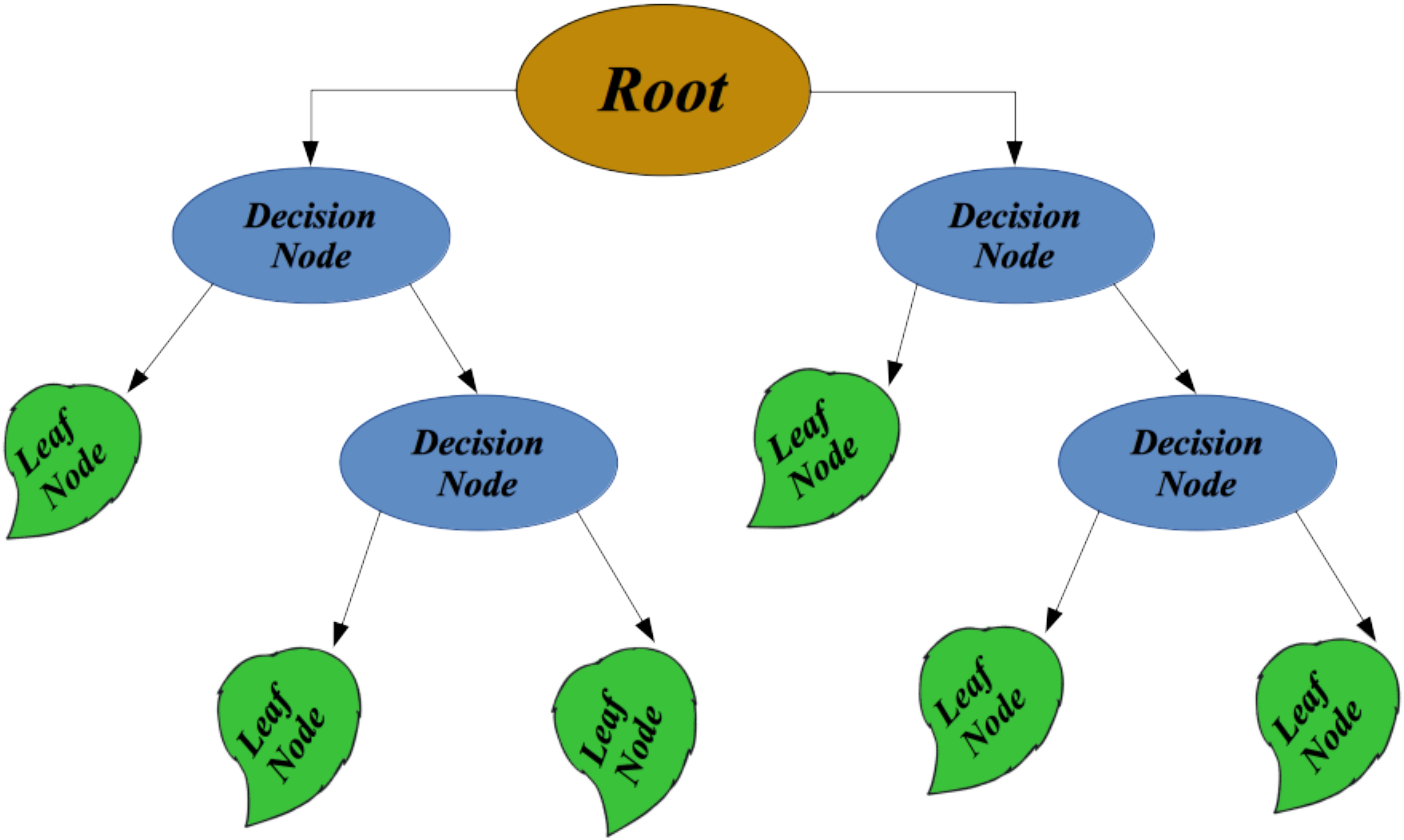
- Supervised algorithm (unsupervised versions exist)
- Hierarchical model
- For Classification and Regression

DECISION TREE STRUCTURE

A Decision Tree is made of :

- A **Root**
- Several **Decision Nodes**
  where tests are performed using a decision rule to split the data set in sub data sets.
- Final **Leaf Nodes**
  at which the output is computed (can be a class or a value)

DECISION TREE STRUCTURE



HOW TO BUILD IT

Ultimately we want the best, simplest and smallest tree possible.

Main issues when building the tree :

- Which test to run at a decision node ? (type, criterion, etc)
- Shall we limit the tree and how ?
- How to estimate the tree quality ?

MAKING DECISION

For a decision rule we can :

- Use a threshold or an interval when decision based on continuous features.
- Use a single feature or several features

UNIVARIATE Vs MULTIVARIATE

- Split the data in two or more

BINARY Vs N-ARY

for instance if classification with ***n*** class, naturally split in ***n*** at each decision node (!careful when ***n*** is large!)

… But how to evaluate the "goodness" of a split ?

GOODNESS OF A SPLIT

Impurity Measure (for classification)

A split is pure if all samples in child nodes belong to the same class

for node *m* , let define the purity as : $p_m^i = \frac{N_m^i}{N_m}$

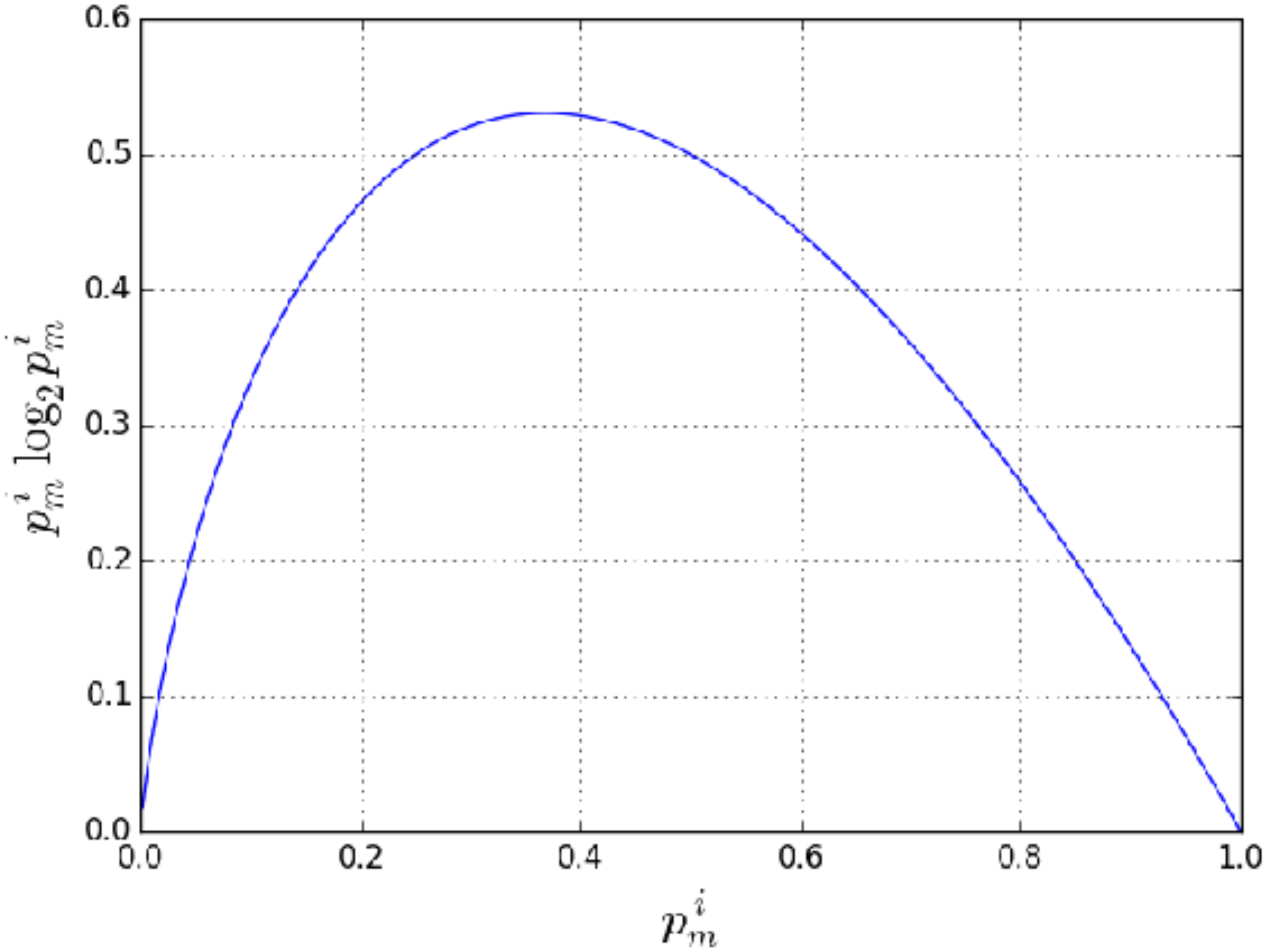
where *i* is the class.

Then a node is pure if : $\forall i \;\; p_m^i = 0 \quad \text{or} \quad p_m^i = 1$

…obviously : $\sum_i^k p_m^i = 1$

ENTROPY FOR DECISION

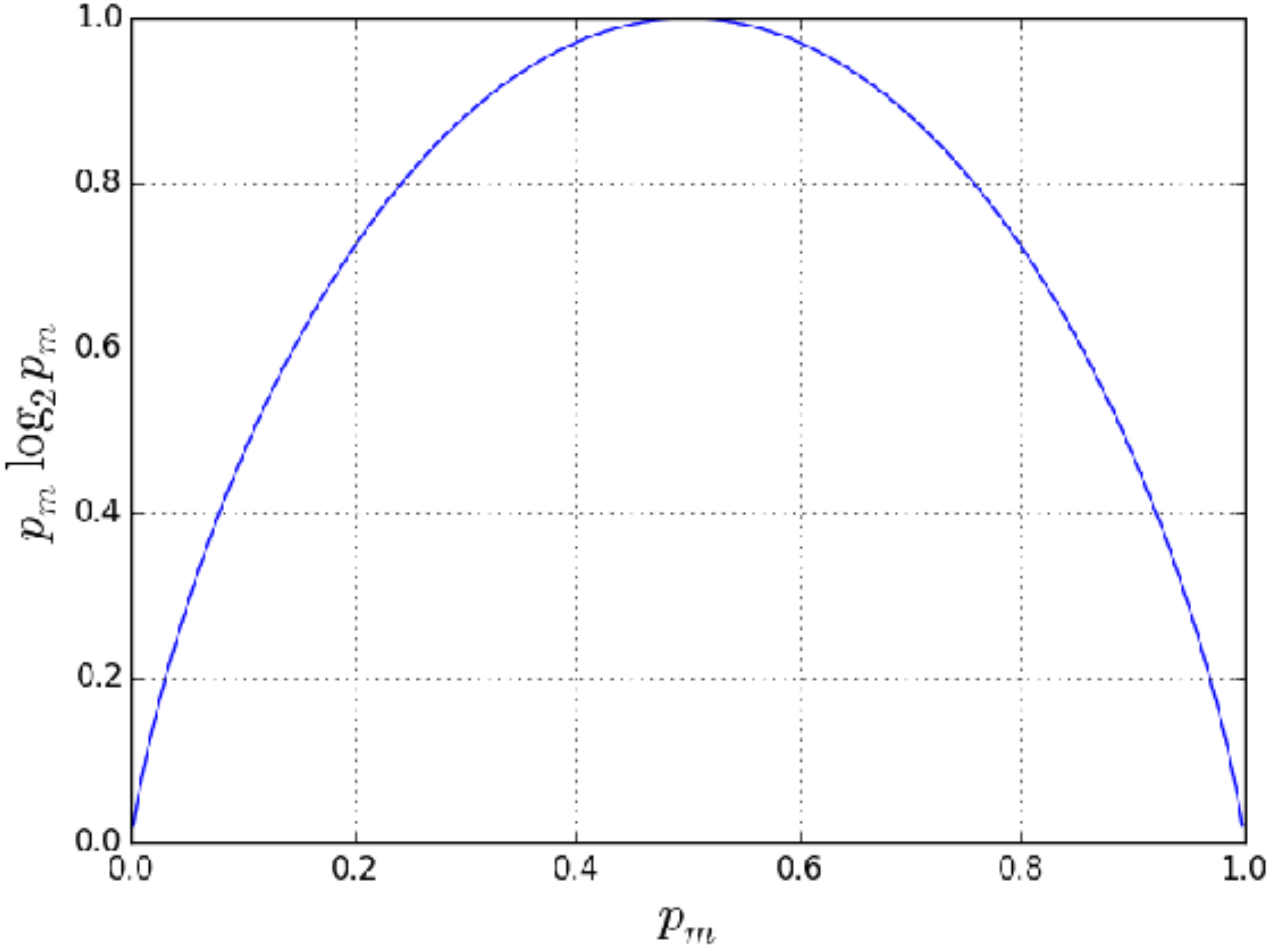
We can minimise the ***entropy*** [Quinlan 1986] : $\Phi_m = -\sum_i^k p_m^i \log_2(p_m^i)$



ENTROPY FOR DECISION

In the case of bimodal split (k=2) : $\Phi(p^1, p^2) = -(p^1 \log_2(p^1) + p^2 \log_2(p^2))$

$$\Phi(1,0) = \Phi(0,1) = 0.0 \quad \text{and} \quad \Phi(0.5, 0.5) = 1.0 \;\; \text{is a maxima}$$



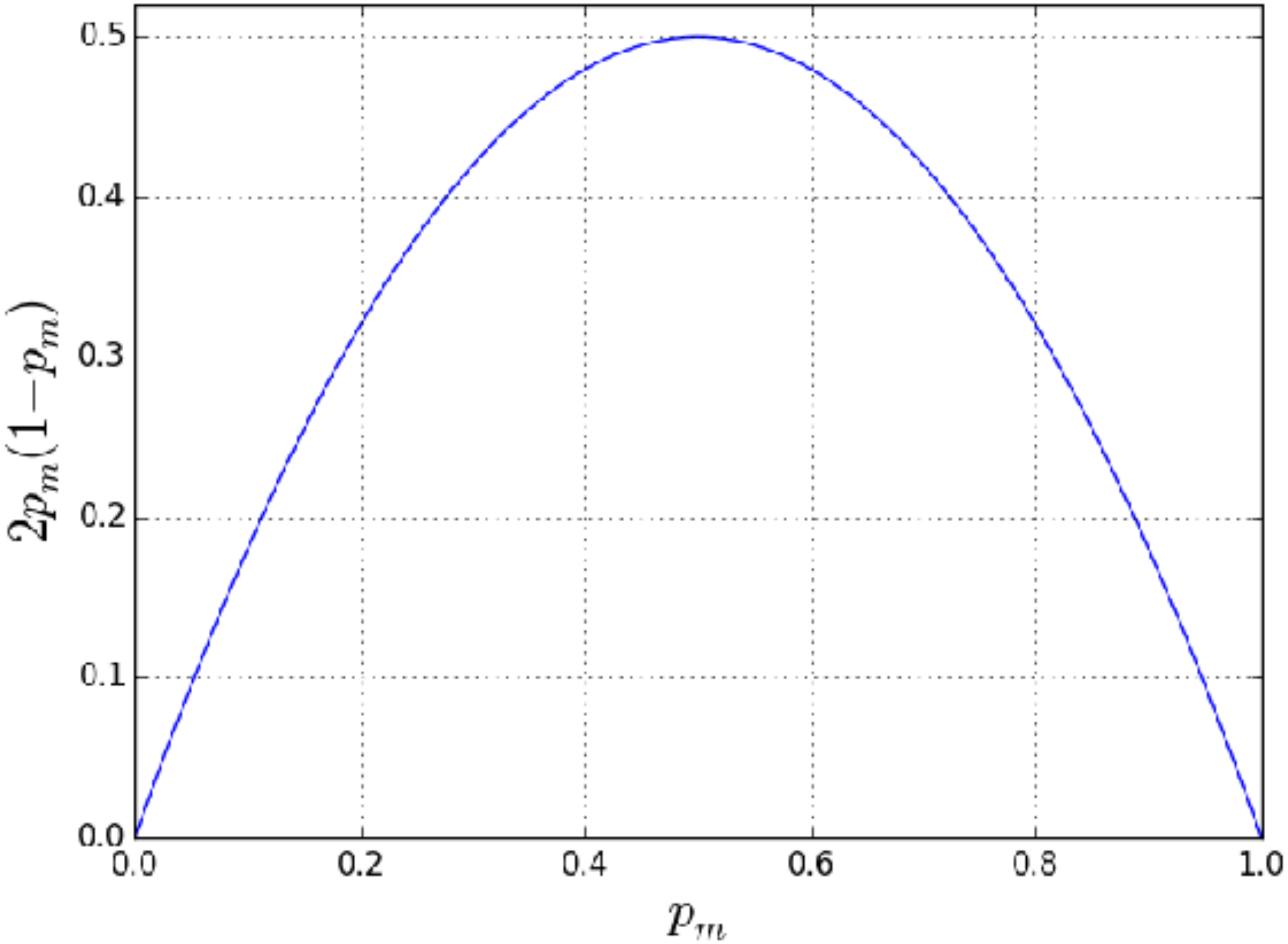
GINI INDEX FOR DECISION

Alternatively minimising the ***Gini index*** : $\Phi_m = \sum^{k} p_m^i (1 - p_m^i)$

Bimodal case :

$$p_m = p_m^1 = 1 - p_m^2$$

$$\Phi_m = 2p_m(1 - p_m)$$



BUILDING THE TREE

- Choose an impurity measure
- At each decision node chose the best feature for splitting i.e. the one that reduce the most impurity
- Construct the tree until all leaf nodes are pure

… which has severe drawbacks!

DRAWBACKS

- Time and resource consuming
- Local optimal decision (at a decision node) does not imply global optimum tree
- Build complex tree and problem of overfitting
- Unstable solution
  (a small change in training set can result in a very different tree)

Setting limits and controlling the construction of the tree helps

LIMITING THE TREE

- Setting a target purity for leaf node
- Limit the depth of the tree
- Stoping when the number of samples reaching a node is inferior to a minimum value or to a fraction of the whole dataset : ***Prepruning***
- Cut useless branches causing overfitting (requires a validation set) : ***Postpruning***

CODING TIME

- Use `DecisionTreeClassifier` to classify the iris dataset.
- Try the two different purity estimators. Does it change the tree?
- Play with the different parameters!