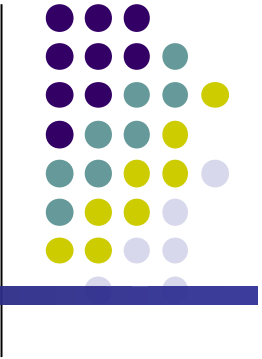
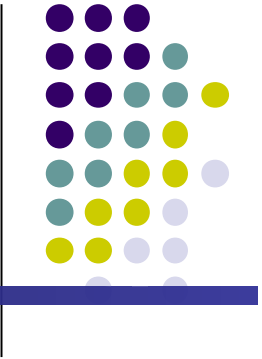
M1 Informatique



Apprentissage Artificiel
Machine Learning

Année 2012-2013 Béatrice Duval

Chapitre 3



Arbres de décision

La tâche d'apprentissage

- X: ensemble de cas décrits par un ensemble d'attributs (nominaux ou continus) {A1,A2, …An}
- c, la fonction cible qui est une fonction de classification
- S: ensemble d'exemples d'apprentissage <x, c(x)>
- H, espace des hypothèses, est constitué par les arbres de décision portant sur les attributs {A1,A2, …An}

On cherche un arbre de H qui pour chaque élément de x de X conduit à une décision h(x) telle que h(x) = c(x), pour ***presque tous les x de S***

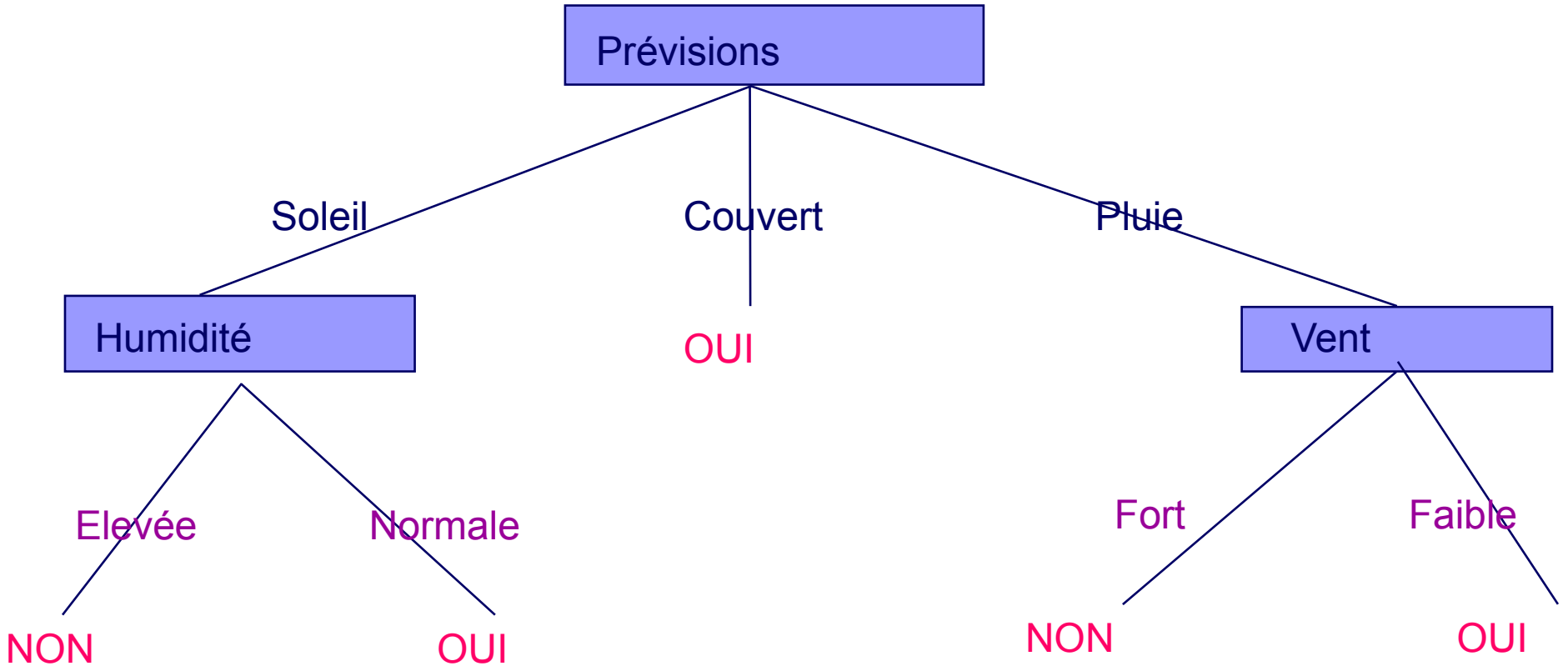
Arbre de décision

Un ensemble de cas E1-E14 décrivant des situations Météo et la décision associée: Sport

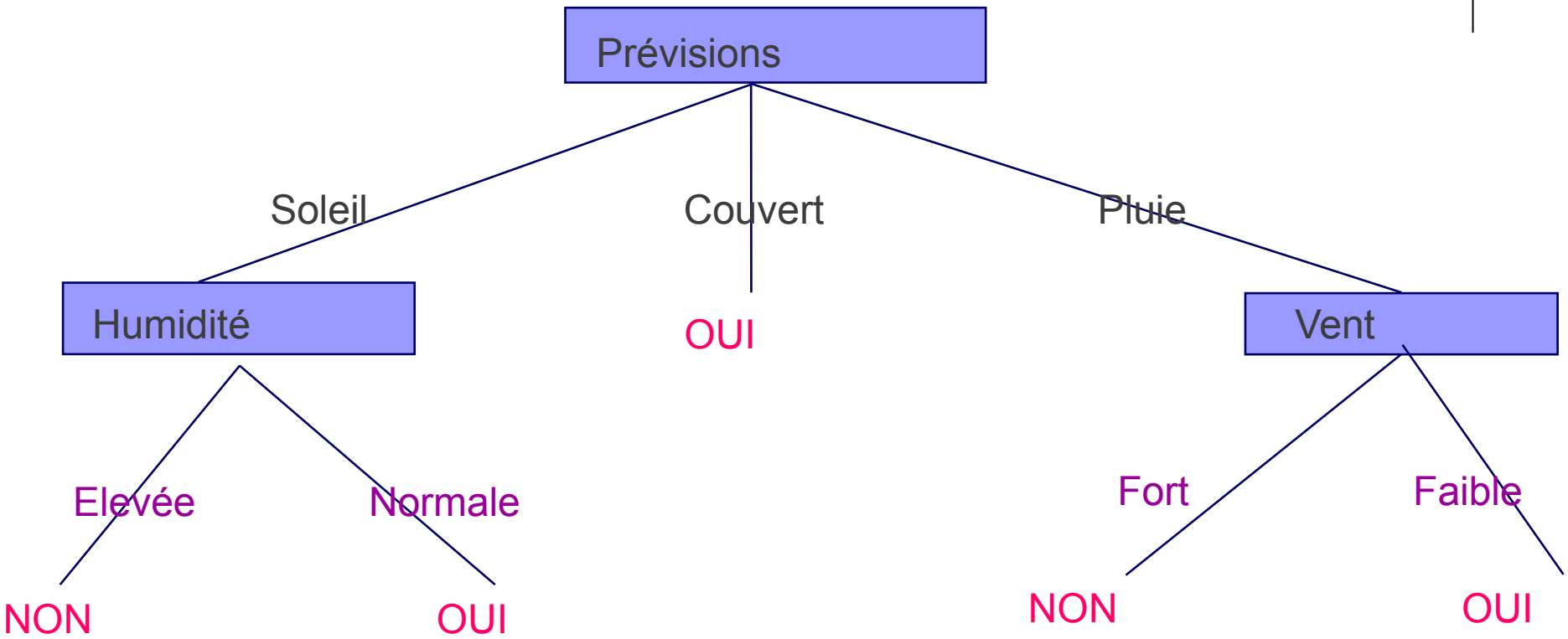
Cas	Prévisions	Température	Humidité	Vent	**Sport**
E1	Soleil	Chaude	Elevée	Faible	NON
E2	Soleil	Chaude	Elevée	Fort	NON
E3	Couvert	Chaude	Elevée	Faible	OUI
E4	Pluie	Douce	Elevée	Faible	OUI
E5	Pluie	Froide	Normale	Faible	OUI
E6	Pluie	Froide	Normale	Fort	NON
E7	Couvert	Froide	Normale	Fort	OUI
E8	Soleil	Douce	Elevée	Faible	NON
E9	Soleil	Froide	Normale	Faible	OUI
E10	Pluie	Douce	Normale	Faible	OUI
E11	Soleil	Douce	Normale	Fort	OUI
E12	Couvert	Douce	Elevée	Fort	OUI
E13	Couvert	Chaude	Normale	Faible	OUI
E14	Pluie	Douce	Elevée	Fort	NON

Construire un arbre qui indique en fonction des valeurs des attributs (Prévisions, Température, Humidité, Vent) quelle décision est prise

Représentation par arbre de décision



Convertir un arbre en règles



SI Prévisions= Soleil et Humidité = Elevée ALORS Sport= NON

SI Prévisions= Soleil et Humidité = Normale ALORS Sport= OUI

SI Prévisions= Couvert ALORS Sport= OUI

SI Prévisions= Pluie et Vent = Fort ALORS Sport= NON

SI Prévisions= Pluie et Vent = Faible ALORS Sport= OUI

Arbre de décision

- Méthode de classification et de prédiction
- Les attributs apparaissant dans l'arbre sont les attributs pertinents pour le problème considéré
- Un arbre est équivalent à un ensemble de règles de décision

Représentation par arbre de décision

- Chaque Nœud interne teste un attribut
- Chaque branche correspond à une valeur d'attribut (cas des attributs nominaux)
- Chaque feuille correspond à une classe unique (décision OUI ou décision NON) ou à une classe majoritaire
- On cherche un arbre le plus « simple » possible expliquant l'ensemble des cas

Construction Top-down d'Arbres de décision

1. *A* ← le «meilleur» attribut de décision pour le nœud courant
2. *Noeud* ← *A* comme attribut
3. Pour chaque valeur de *A*, créer un nouveau descendant de *Noeud*

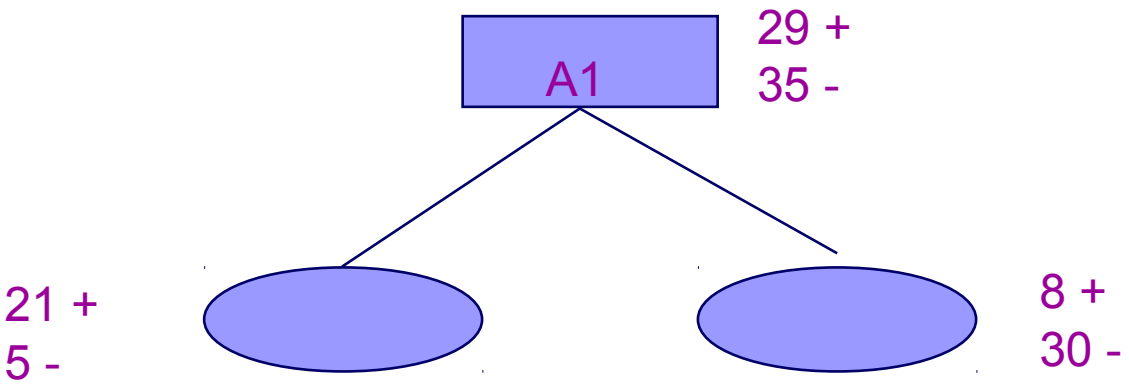
4 Trier les exemples dans les feuilles

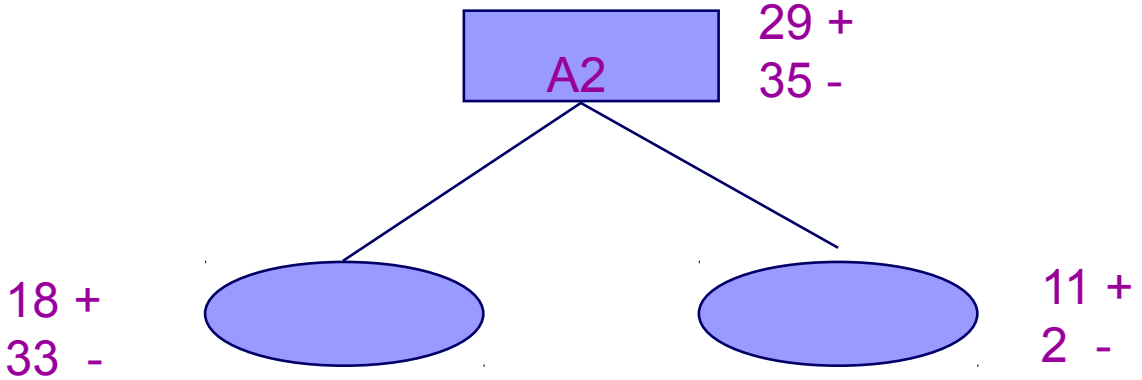
5. Si chaque feuille satisfait le critère d'arrêt

   alors Fin

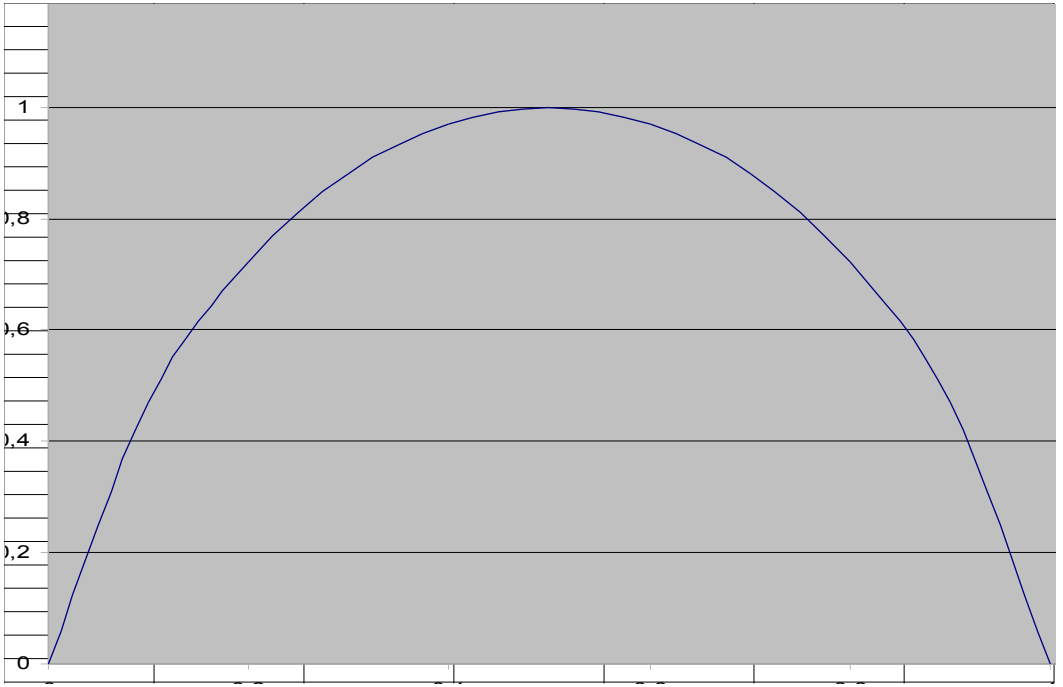
   Sinon Itérer sur les feuilles

Choix du meilleur attribut





Entropie



- S est un ensemble d'exemples
- p_+ proportion d'ex. positifs dans S
- p_- proportion d'ex. négatifs dans S
- L'entropie mesure l'impureté de S
- $Ent(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$

Entropie

- Mesure issue de la théorie de l'information
- Entropie(S) = Nombre de bits nécessaire pour coder la classe(+ ou -) d'un élément tiré au hasard

Gain d'information (C4.5)

- Gain(S, A)= réduction d'entropie due au test de l'attribut A

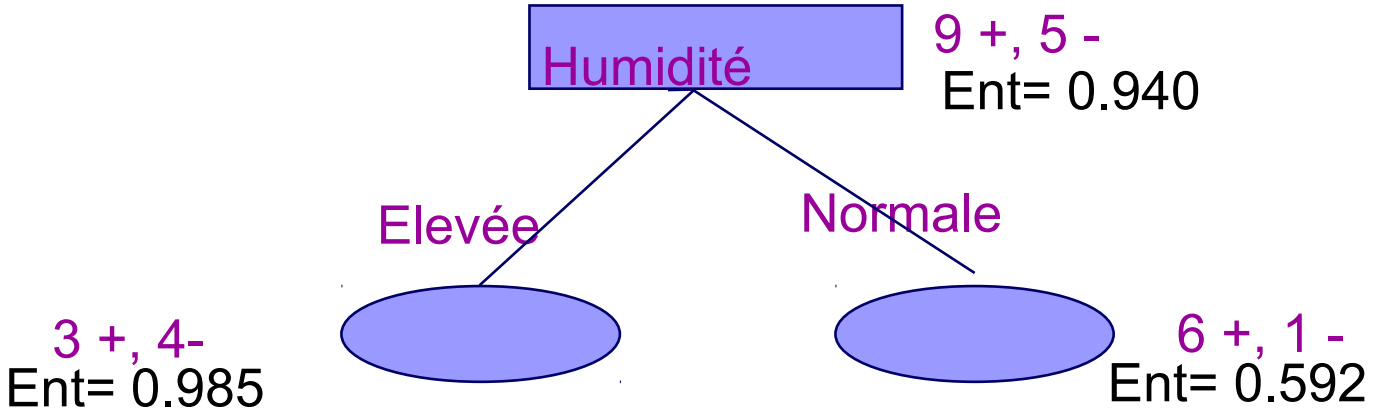
$$Gain(S, A) = Ent(S) - \sum_{v \in Valeurs(A)} \frac{|S_v|}{|S|} Ent(S_v)$$

- Si on travaille avec plus de 2 classes, la formule d 'entropie peut être généralisée

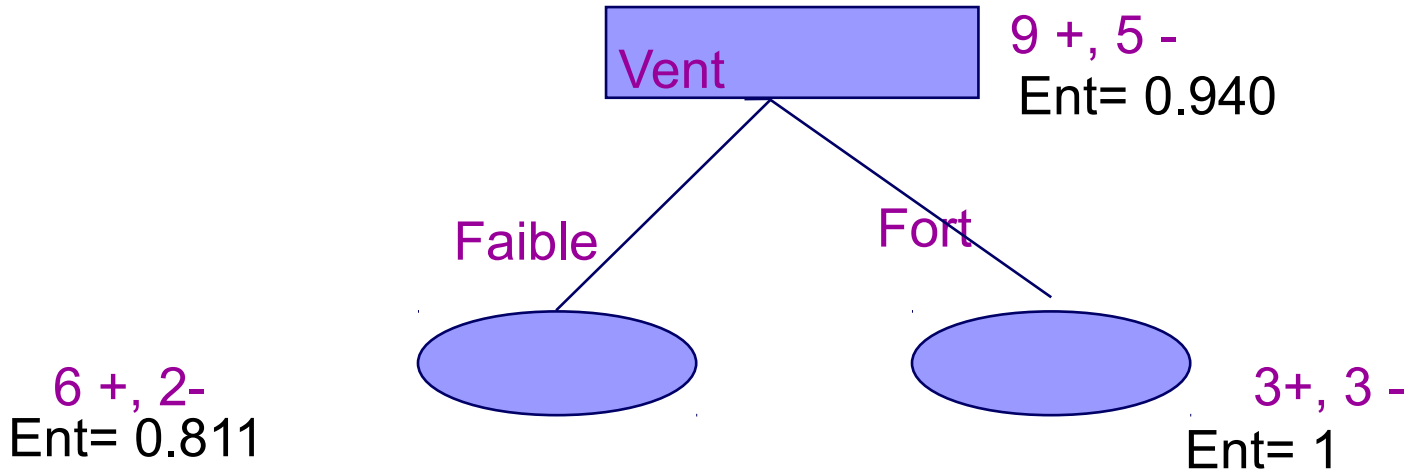
$$Ent(S) = -\sum_{s} f_s \log_2 f_s$$

où f_s est la proportion de la classe s

Gain d ’information

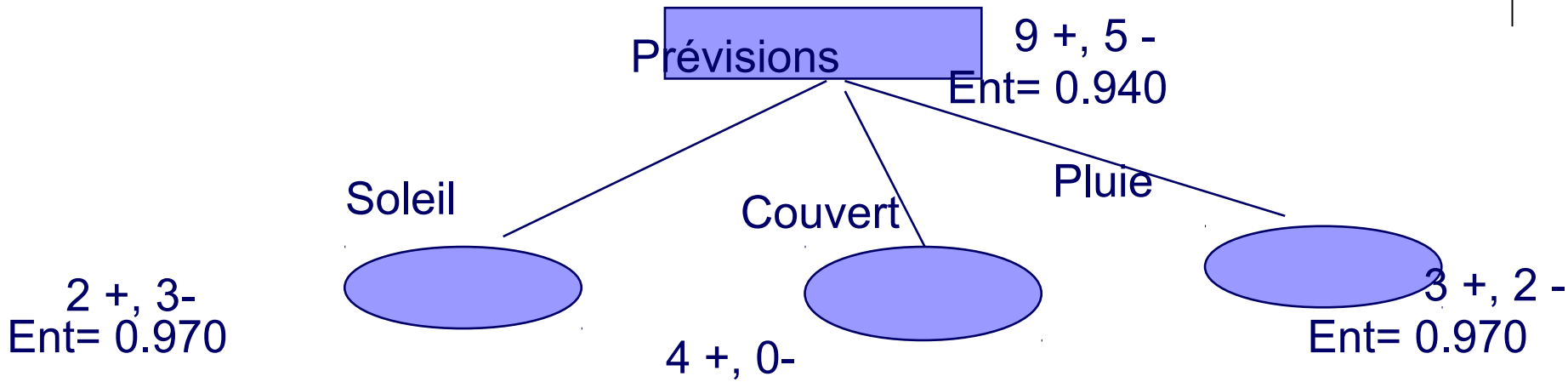


Gain(S, Humidité)=0.940 - 7/14*0.985 - 7/14*0.592
Gain(S, Humidité)=0.151



Gain(S, Vent)=0.940-8/14*0.811 -6/14*1. = 0.048

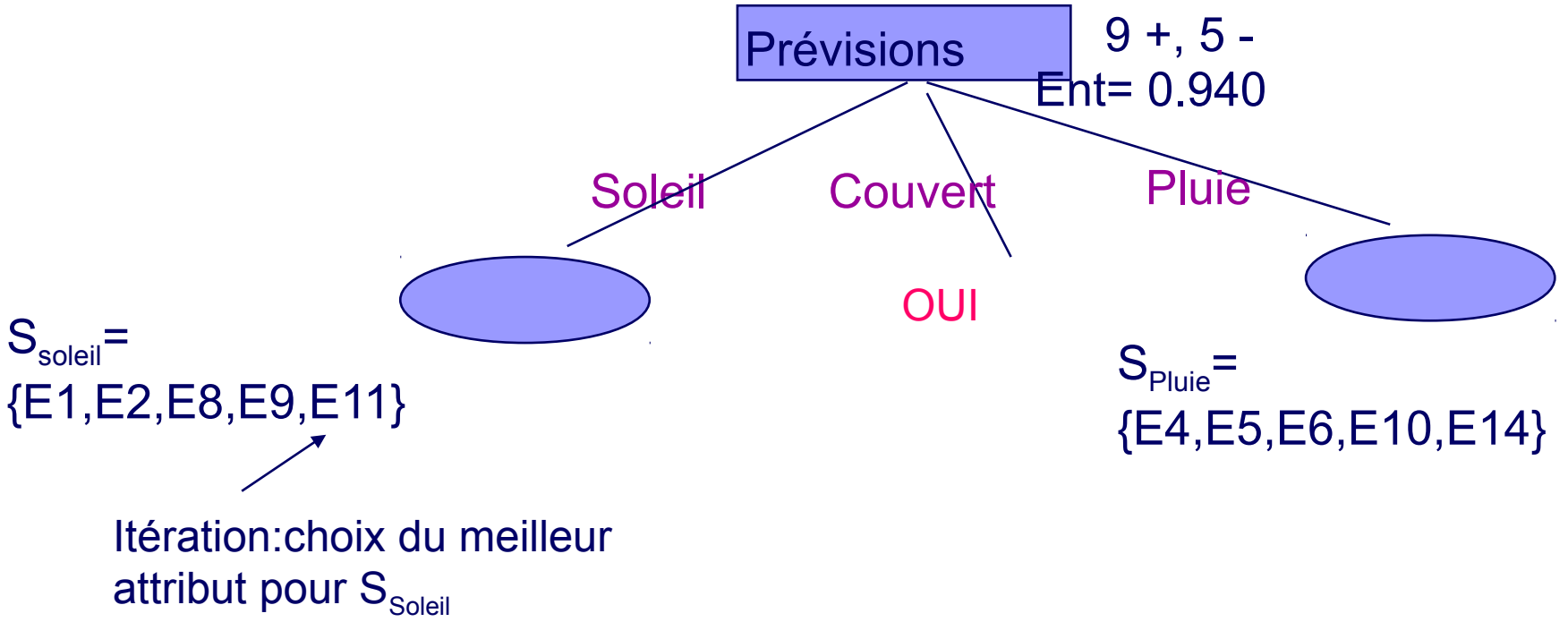
Gain d ’information



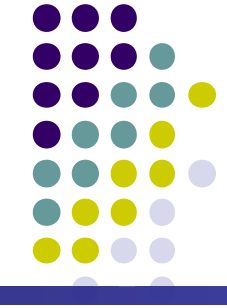
Gain(S, Prévisions)=0.940 - 5/14*0.970 - 5/14*0.970 = 0.246

Gain(S, Temp)=0.940 - 4/14 - 6/14*0.918 – 4/14*0.811= 0.029

Construction de l'arbre



Attributs à valeurs continues



- Pour un attribut A continu, créer dynamiquement un attribut binaire à partir des valeurs observées, dans le nœud considéré

- On trie les valeurs observées pour A
  - On cherche un seuil s tq l'attribut discrétisé

    (A≤ s ou A >s)

  donne le meilleur gain d 'information

Surajustement de l'arbre de décision

- Surajustement= arbre trop « proche » des données
- Arbre peu performant pour la prédiction
- Le bruit dans les exemples peut conduire à un surajustement de l 'arbre

  E15= (Soleil,Chaude,Normale, Fort, NON)

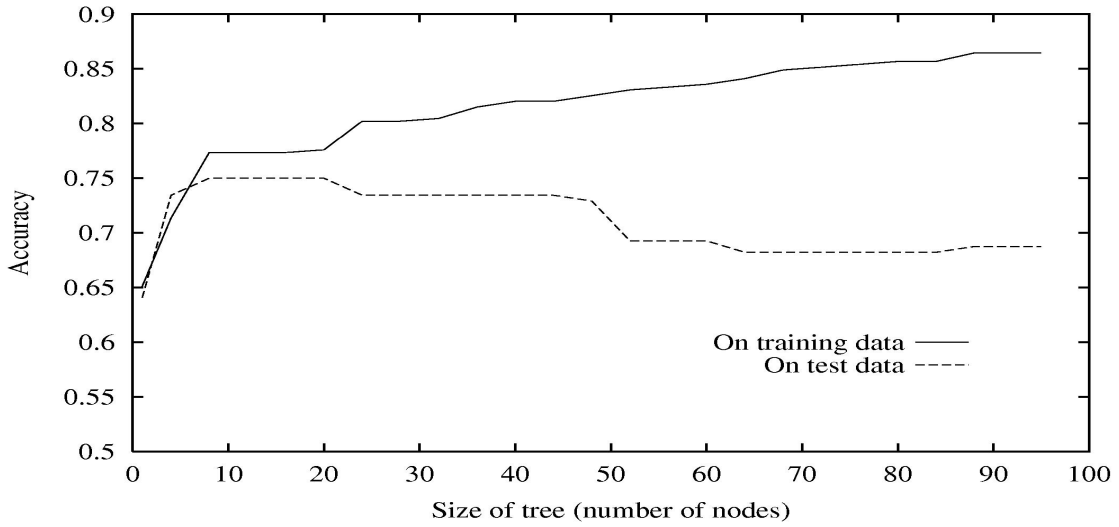
  Conséquences pour l 'arbre construit?

Eviter le surajustement

- Comment éviter le surajustement?
  - Construire l ’arbre complet puis l’élaguer: estimation de l'arbre élagué sur un ensemble test
  - Transformer l ’arbre en ensemble de règles et simplifier les règles

Estimation de l'erreur

- Si on dispose de beaucoup de données étiquetées, on les sépare en EA: ensemble d ’apprentissage et ET: ensemble de test on s’assure que chaque classe est suffisamment représentée dans EA et ET
- L'erreur du classifieur est mesurée par le taux de mal classés
- Le taux d'erreur du classifieur construit à partir de EA est évaluée sur ET, l'ensemble de test


0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
0
10
20
30
40
50
60
70
80
90
100
Size of tree (number of nodes)
On training data
On test data

Coût de l ’erreur en classification

- 2 classes OUI/NON
- Matrice de confusion

	Prédit OUI	Prédit NON
Classe réelle OUI	Vrai positif	Faux négatif
Classe réelle NON	Faux positif	Vrai négatif