Les arbres de décisions

Construction et évaluation

- Quels attributs sélectionner pour construire l'arbre ?
- Comment évaluer un arbre ? A priori on souhaite minimiser le nombre de tests (en moyenne) pour classifier un exemple (il existe une borne max sur le nombre d'attributs).

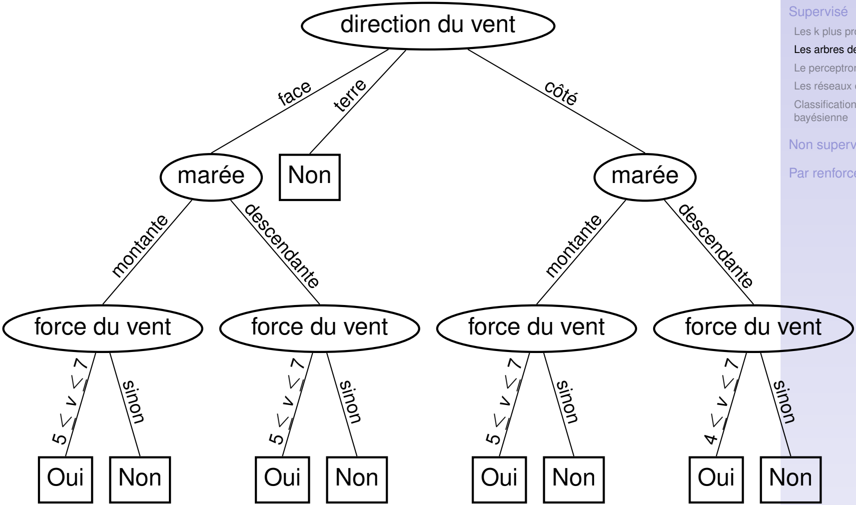
Les arbres de décisions

Un premier exemple

- La force du vent : attribut numérique {1 ...12}.
- La direction du vent : attribut nominal {face, terre, côté}
- La marée : attribut nominal {montante, descendante}

- Décision à prendre : puis-je faire de la planche à voile ?

Les arbres de décisions



Les arbres de décisions

Comment construire un arbre de décisions

- Parcours exhaustif des arbres impossible.
- Exponentiel en fonction de
  - nombre d'attributs d.
  - nombre moyen de valeurs par attribut : v.
  - $\sum_{i=0}^{d-1} (d-i)^{v^i}$
  - Par exemple, avec 6 attributs et $v = 2$ on a déjà 72385.

Construction intelligente

Algorithm 1 Construction récursive d'un arbre de décision

```
Procédure : construireArbre(X)
if Tous les points de X appartiennent à la même classe then
    Créer une feuille portant le nom de cette classe
else
    Choisir le meilleur attribut pour créer un noeud
    Le test associé à ce noeud sépare X en X_g et X_d.
    construireArbre(X_g)
    construireArbre(X_d)
end if
```

Une interprétation probabiliste

Quelques définitions

- Un exemple n_i est décrit par :
  - d attributs.
  - u une classe $\in U = \{u_1, \ldots, u_c\}$
- Soient n points de l'échantillon d'apprentissage, répartis en U classes u_j comportant chacune n_j exemples.
- Soit a un attribut binaire quelconque.
- a partage chaque n_j en deux sous parties, comportant :
  - l_j points pour $a = vrai$.
  - r_j points pour $a = faux$.

Une interprétation probabiliste

On peut alors déduire que

- l_j/n est une estimation de $P(a = vrai, u = u_j)$.
- r_j/n est une estimation de $P(a = faux, u = u_j)$.

- l/n est une estimation de $P(a = vrai)$.
- r/n est une estimation de $P(a = faux)$.

- n_j/n est une estimation de $P(u = u_j)$.

Une mesure pour choisir l'attribut

Théorie de l'information

- Information mutuelle (entropie croisée).
- Permet de mesurer l'homogénéité entre deux distributions de probabilités.

- Soient u et a deux variables, avec D_u et D_a leurs ensembles finis de valeurs possibles.
- L'entropie croisée est donnée par :

$$\mathtt{I}(u,a) = -\sum_{i,j \in D_u x D_a} p(i,j) \log_2 \frac{p(i,j)}{p(i)p(j)}$$

Une mesure pour choisir l'attribut

Quelques propriétés

- $\mathtt{I}(u, a)$ possède un minimum (0) quand $p(i, j) = p(i)p(j)$: lorsque les deux distributions sont indépendantes.
- $\mathtt{I}(u, a)$ est maximale lorsque les deux distributions sont complètement corrélées.

Une mesure pour choisir l'attribut

Quelques propriétés

- Une variable aléatoire w possède une entropie $H(u)$ définit par :

$$H(u) = -\sum_{i \in D_u} p(i) \log_2(p(i))$$

- L'entropie de w conditionnée par a est définie par :

$$H(u|a) = -\sum_{i,j \in D_u x D_a} p(i,j) \log_2(p(i|j))$$

- Or d'après la théorie de l'information, on a :

$$\mathtt{I}(u,a) = H(u) - H(u|a)$$

Une mesure pour choisir l'attribut

- $\hat{\imath}(u,a) = -\sum_{j=1}^{Y} \frac{l_j}{n} \log_2 \frac{l_j/n}{(l/n)(n_j/n)} + \frac{r_j}{n} \log_2 \frac{r_j/n}{(r/n)(n_j/n)}$
- $\hat{H}(u|a) = -\sum_{j=1}^{Y} \frac{l}{n}\frac{l_j}{l} \log_2 \frac{l_j}{l} + \frac{r}{n}\frac{r_j}{r} \log_2 \frac{r_j}{r}$

Une mesure pour choisir l'attribut

Quelques propriétés

- On peut donc estimer les probabilités :

$$\hat{H}(u|a) = \frac{l}{n} J(a = vrai) + \frac{r}{n} J(a = faux)$$

avec

$$J(a = vrai) = -\sum_{j=1}^{U} \frac{l_j}{l} \log_2 \frac{l_j}{l}$$

$$J(a = faux) = -\sum_{j=1}^{U} \frac{r_j}{r} \log_2 \frac{r_j}{r}$$

Une mesure pour choisir l'attribut

Choix de l'attribut

- On veut choisir l'attribut qui possède la plus grande corrélation avec la répartition des classes.
- On va donc chercher à minimiser l'entropie :

$$i^* = \arg\min_{i=1,\ldots,d}(\hat{H}(u|a_i))$$

Un exemple de construction

Problème

Le problème consiste à prédire si un enfant peut aller jouer avec son voisin.

Base d'exemples

Les décisions prises sur les 8 derniers jours étant donnés 4 attributs binaires et 1 classe.

Un exemple de construction

Échantillon d'apprentissage

	Devoirs finis ?	Bonne humeur de la mère ?	Beau temps ?	Gouter pris ?	**Décision**
1	vrai	faux	vrai	faux	**oui**
2	faux	vrai	faux	vrai	**oui**
3	vrai	vrai	vrai	faux	**oui**
4	vrai	faux	vrai	vrai	**oui**
5	faux	vrai	vrai	vrai	**non**
6	faux	vrai	faux	faux	**non**
7	vrai	faux	faux	vrai	**non**
8	vrai	vrai	faux	faux	**non**

Simplifions les notations : (DF, BH, BT, GP).

Un exemple de construction

	Devoirs finis ?	Bonne humeur de la mère ?	Beau temps ?	Gouter pris ?	**Décision**
1	vrai	faux	vrai	faux	**oui**
2	faux	vrai	faux	vrai	**oui**
3	vrai	vrai	vrai	faux	**oui**
4	vrai	faux	vrai	vrai	**oui**
5	faux	vrai	vrai	vrai	**non**
6	faux	vrai	faux	faux	**non**
7	vrai	faux	faux	vrai	**non**
8	vrai	vrai	faux	faux	**non**

Calcul de $H(u = oui|DF)$

$$H(oui|DF) = \frac{5}{8} J(DF = vrai) + \frac{3}{8} J(DF = faux)$$

$$J(DF = vrai) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$J(DF = faux) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}$$

Un exemple de construction

On obtient

- $H(oui|DF) \approx 0.9$
- $H(oui|BT) \approx 0.8$
- $H(oui|BH) \approx 0.9$
- $H(oui|GP) \approx 1$

On choisit donc pour racine l'attribut “Est ce qu'il fait beau ?”

Un exemple de construction

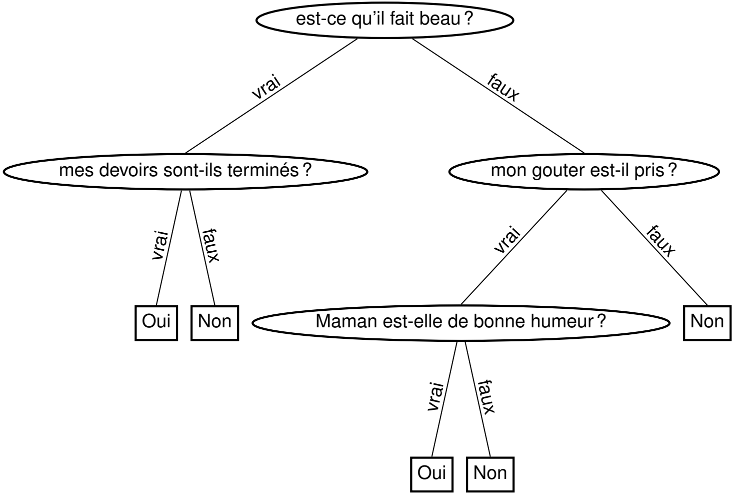
Table pour la valeur vrai

	Devoirs finis ?	Bonne humeur de la mère ?	Gouter pris ?	**Décision**
1	vrai	faux	faux	**oui**
3	vrai	vrai	faux	**oui**
4	vrai	faux	vrai	**oui**
5	faux	vrai	vrai	**non**

Table pour la valeur faux

	Devoirs finis ?	Bonne humeur de la mère ?	Gouter pris ?	**Décision**
2	faux	vrai	vrai	**oui**
6	faux	vrai	faux	**non**
7	vrai	faux	vrai	**non**
8	vrai	vrai	faux	**non**

Un exemple de construction



Élagage des arbres de décisions

Pourquoi élaguer ?

- L'arbre précédemment construit est dit T_{max} : chaque feuille est pure.
- Il y a un risque de sous estimer la probabilité d'erreur.

Le pré-élagage

- On cesse de diviser un noeud lorsque la pureté des points est suffisantes (inférieure à un certain seuil).
- Utilisation de critères locaux (à une feuille) : on peut manquer un très bon développement.
- Donc, utilisation du post-élégage.

Élagage des arbres de décisions

Le post-élagage

- Élaguer l'arbre lorsqu'il est parfaitement développé.
- Utiliser un ensemble indépendant de l'ensemble d'apprentissage (ensemble de validation).
- Mesurer l'erreur commise sur cet ensemble.

Méthode

- En partant des feuilles, construire une séquence d'arbres $\{T_{max}, T_1, T_2, \ldots, T_n\}$ (T_n est l'arbre constitué d'une seule feuille).
- A chaque étape un noeud est transformé en feuille.
- L'idée est de comparer le coût de l'arbre élagué et de l'arbre non élagué. On s'arrête si le coût du premier est supérieur.

Élagage des arbres de décisions

Estimation du coût

Choisir le noeud v qui minimise :

$$\hat{w}(T_k, v) = \frac{MC_{ela}(v,k) - MC(v,k)}{n_k.(nt(v,k) - 1)}$$

- $MC_{ela}(v,k)$: nombre d'exemples de l'ensemble d'apprentissage mal classés **par** le noeud v de T_k dans l'arbre élagué à v.
- $MC(v,k)$: nombre d'exemples de l'ensemble d'apprentissage mal classés **sous** le noeud v de T_k dans l'arbre **non** élagué à v.
- n_k : nombre de feuilles de T_k.
- $nt(v,k)$: nombre de feuilles du sous-arbre de T_k situé sous le noeud v.

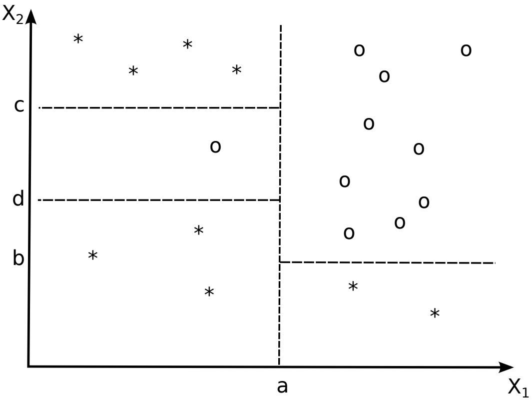
On cherche le meilleur compromis entre le taux d'erreur apparent et la taille.

Élagage d'un arbre de décisions

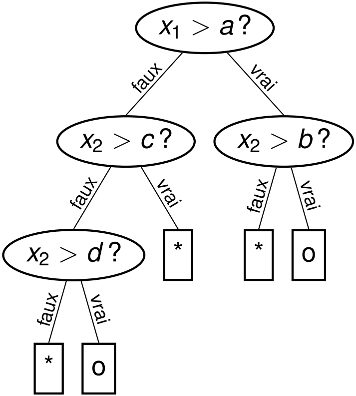
Algorithm 2 Algorithme d'élagage

Procédure : `elaguer`(T_{max})
$k \leftarrow 0$
$T_k \leftarrow T_{max}$
while `nbNoeud`$(T_k) > 1$ **do**
    for chaque noeud v de T_km **do**
        calculer $\hat{w}(T_k, v)$ sur l'ensemble d'apprentissage
    end for
    Choisir le noeud v_m pour lequel $\hat{w}$ est minimum
    T_{k+1} se déduit de T_k en y remplaçant v_m par une feuille
    $k \leftarrow k + 1$
end while

Un exemple d'élagage

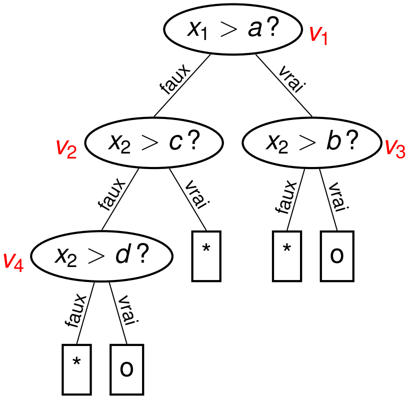


Un exemple d'élagage



Un exemple d'élagage

Calculer $\hat{w}(T_{max}, v_1)$, $\hat{w}(T_{max}, v_2)$, $\hat{w}(T_{max}, v_3)$, $\hat{w}(T_{max}, v_4)$



Un exemple d'élagage

Calcul des valeurs sur T_{max}

- $\hat{w}(T_{max}, v_1) = \dfrac{MC_{ela}(v_1, k) - MC(v_1, k)}{n(k).(nt(v_1, k) - 1)} = \dfrac{9 - 0}{5(5 - 1)} = \dfrac{9}{20}$
- $\hat{w}(T_{max}, v_2) = \dfrac{1 - 0}{5(3 - 1)} = \dfrac{1}{10}$
- $\hat{w}(T_{max}, v_3) = \dfrac{2 - 0}{5(2 - 1)} = \dfrac{2}{5}$
- $\hat{w}(T_{max}, v_4) = \dfrac{1 - 0}{5(2 - 1)} = \dfrac{1}{5}$

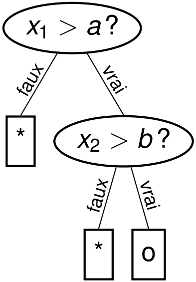
Un exemple d'élagage

Calcul des valeurs sur T_{max}

- $\hat{w}(T_{max}, v_1) = \dfrac{MC_{ela}(v_1, k) - MC(v_1, k)}{n(k).(nt(v_1, k) - 1)} = \dfrac{9-0}{5(5-1)} = \dfrac{9}{20}$
- $\hat{w}(T_{max}, v_2) = \dfrac{1-0}{5(3-1)} = \dfrac{1}{10}$
- $\hat{w}(T_{max}, v_3) = \dfrac{2-0}{5(2-1)} = \dfrac{2}{5}$
- $\hat{w}(T_{max}, v_4) = \dfrac{1-0}{5(2-1)} = \dfrac{1}{5}$

Un exemple d'élagage

On obtient T_1



Un exemple d'élagage

Calcul des valeurs sur T_1

- $\hat{w}(T_1, v_1) = \frac{9-1}{3(3-1)} = \frac{4}{3}$
- $\hat{w}(T_1, v_3) = \frac{2-1}{3(2-1)} = \frac{1}{3}$

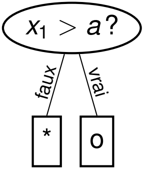
Un exemple d'élagage

Calcul des valeurs sur T_1

- $\hat{w}(T_1, v_1) = \frac{9-1}{3(3-1)} = \frac{4}{3}$
- $\hat{w}(T_1, v_3) = \frac{2-1}{3(2-1)} = \frac{1}{3}$

Un exemple d'élagage

On obtient T_2



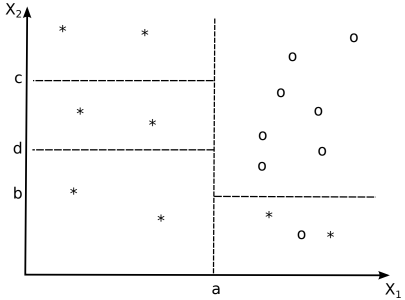
Un exemple d'élagage

Choix de l'arbre

- Le choix de l'arbre se fera sur un ensemble de validation parmi T_{max}, T_1 et T_2.
- L'arbre qui aura la meilleure estimation du taux d'erreur de classification sera choisi.

Un exemple d'élagage

Par exemple, sur cet ensemble de validation



- Erreur T_{max} : $\frac{3}{16}$
- Erreur T_1 : $\frac{1}{16}$
- Erreur T_2 : $\frac{2}{16}$