

Sujet d'examen du master ISF

Arbre de décision

On considère le tableau suivant décrivant un arbre de décision obtenu par une méthode de segmentation sur un ensemble d'apprentissage de 80 exemples appartenant à deux classes a priori $\{G_1, G_2\}$. La validation de cet arbre se réalise avec un l'ensemble test de 50 exemples.

La première colonne du tableau représente le numéro du nœud, la seconde le numéro du père et les deux suivantes les effectifs sur l'ensemble d'apprentissage et les deux dernières sur l'ensemble test dans chaque des classes a priori.

| ID | père | App G_1 | App G_2 | Test G_1 | Test G_2 |
|----|------|-----------|-----------|------------|------------|
| 1 | 0 | 50 | 30 | 30 | 20 |
| 2 | 1 | 45 | 10 | 25 | 10 |
| 3 | 1 | 5 | 20 | 5 | 10 |
| 4 | 2 | 10 | 2 | 5 | 1 |
| 5 | 2 | 35 | 8 | 20 | 9 |
| 6 | 5 | 3 | 6 | 4 | 5 |
| 7 | 5 | 32 | 2 | 16 | 4 |
| 8 | 7 | 31 | 0 | 12 | 2 |
| 9 | 7 | 1 | 2 | 4 | 2 |

On se place dans l'hypothèse que les coûts de mauvaise classification sont unitaires et que les probabilités des classes a priori sont égales aux fréquences observées dans cet échantillon.

- 1) Dessiner l'arbre de décision, construire la représentation de cet arbre sous la forme (T, g, d) où $g(t)$ est le numéro du fils gauche de t et $d(t)$ le numéro du fils droit. Donner \tilde{T} qui est l'ensemble des feuilles de cet arbre T .
- 2) Construire une colonne supplémentaire contenant le nom de la classe d'affectation de chaque segment.
- 3) Donner la formule de $r(t)$ qui est la probabilité de l'erreur de classement du segment t et ajouter une colonne contenant cette valeur.
- 4) Calculer la probabilité d'erreur de classement sur l'ensemble d'apprentissage et sur l'ensemble test de l'arbre $R(T) = \sum_{t \in \tilde{T}} p(t)r(t)$ où $p(t)$ est la probabilité qu'un exemple soit mis dans la feuille t .
- 5) Calculer ces probabilités d'erreur de classement des ensembles d'apprentissage et de test pour tous les arbres élagués (regroupez les feuilles 8 et 9, puis les feuilles 6 et 7, puis les feuilles 4 et 5 et enfin les feuilles 2 et 3). Construire l'évolution des courbes pour l'ensemble d'apprentissage et l'ensemble test en fonction du nombre de feuilles dans l'arbre. Comment choisir le « meilleur » arbre ?
- 6) Si le segment terminal t de l'arbre T est divisé en deux segments t_g et t_d alors cet nouvel arbre T' constitué à partir de T vérifie la propriété $R(T') \leq R(T)$.

La question 6 est indépendante des autres questions.