Objectif

Montrer comment, à l'aide de TANAGRA, comparer deux algorithmes d'apprentissage supervisé en évaluant leurs performances à l'aide de la validation croisée.

Les méthodes que l'on veut évaluer sont :

- Méthode des plus proches voisins, celle qui est implémenté utilise une distance, la HVDM (voir référence sur la page WEB), qui lui permet d'appréhender autant les variables discrètes que continues ;
- Méthode d'arbre de décision, très simpliste, dérivée de ID3.

Fichier

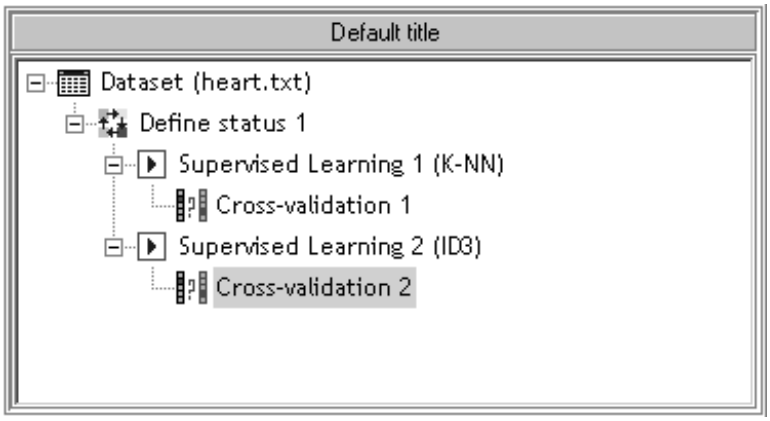
Fichier HEART DISEASES DIAGNOSTIC en provenance de UCI, il y 4 bases disponibles, seule celle en provenance de l'hôpital de Cleveland est utilisée par la communauté apprentissage. La variable à prédire est l'occurrence ou non d'une maladie cardio-vasculaire chez des patients ; les descripteurs correspondent à leurs caractéristiques (sexe, taux de cholestérol, etc.). La variable THAL, qui s'impose comme la meilleure variable dans la prédiction, et dont on ne connaît pas très bien la signification, sur la distribution UCI en tous les cas, a été soustraite du fichier.

Etapes de l'expérimentation

1. Charger le fichier de données
2. Introduire un « Define Status » pour définir le statut des variables :

A prédire	Cœur
Prédictives	Age, Sexe, Type_Douleur, Pression, Cholesterol, Sucre, Electro, Taux_Max, Angine, Depression, Pic, Vaisseau

3. Insérer les deux méthodes supervisées K-NN et ID3. Pour chacune, il faut dans un premier temps placer apprentissage simple (palette Meta-Spv learning) dans laquelle on intègre la méthode supervisée (palette Spvlearning). Introduisez les paramètres ci-dessous pour chaque méthode.
   - K-NN : Nombre de voisins = 5
   - ID3 : Min Size For Split = 20, Min Size of Leaves = 5, Max depth of tree = 10, Min Entropy gain for splitting = 0.03 (*on remarquera au passage que le paramétrage de ID3 est particulièrement difficile et très dépendant du fichier*).
4. Placer, à la suite de chaque apprentissage, un composant validation croisée (palette Spv learning assessment). Laissez les paramètres par défaut. A ce stade, le diagramme de traitement doit avoir l'aspect suivant.



5. Reste alors à exécuter de manière indépendante les deux apprentissages supervisés sur la totalité du fichier, on remarque que les deux méthodes semblent proposer, en re-substitution, les mêmes performances, en termes de taux d'erreur : 0.144 pour le 5-NN, et 0.152 pour ID3. Elles se valent donc apparemment.
6. Pour bien vérifier cette assertion, essayons maintenant d'évaluer l'erreur en validation croisée, 5 x 2-CV tel qu'il est préconisé par DIETTERICH. Les erreurs affichées se démarquent nettement cette fois-ci, sur cette base, la méthode 5-NN apparaît nettement plus performante finalement : **#0.19 pour le 5-NN** et **#0.26 pour ID3** ; à comparer avec le taux d'erreur du **classifieur par défaut (#0.44)** qui préconise de prédire systématiquement « absence de maladie ».
7. Moralité : ne jamais utiliser les mêmes individus pour construire et évaluer les méthodes d'apprentissage supervisé.