Objectif

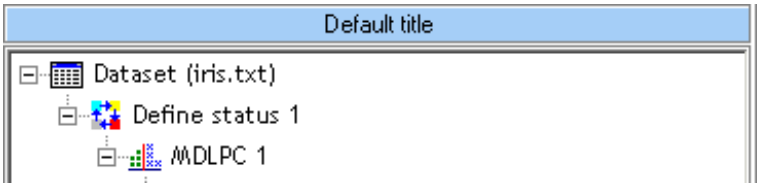
Montrer l'influence de la sélection de variables sur les performances du modèle bayesien naïf.

Fichier

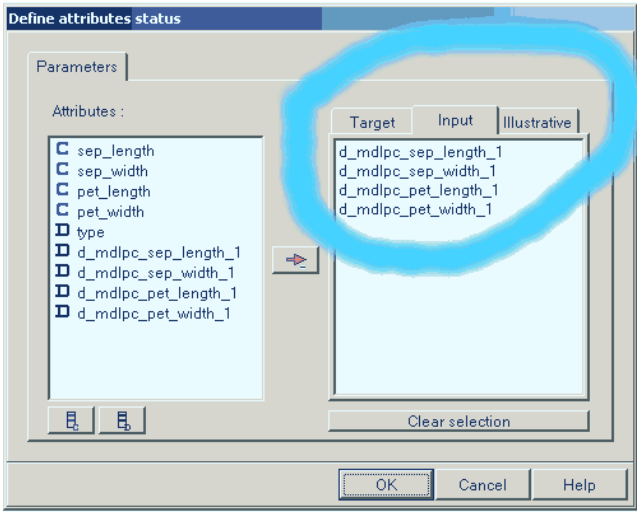
Le fichier IRIS de Fisher, très utilisé en apprentissage automatique, son principal intérêt est que l'on connaît à l'avance le « bon » résultat.

Sélection de variables pour l'apprentissage supervisé

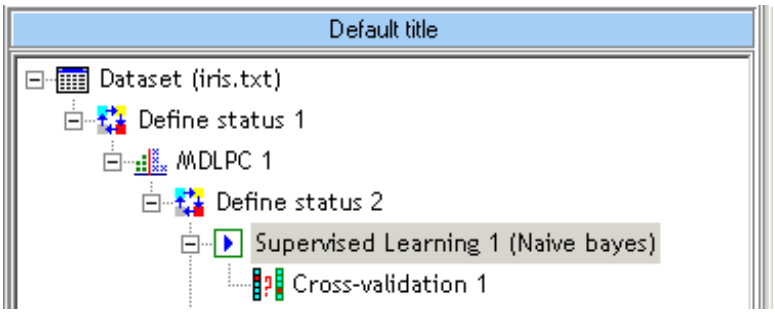
1. Charger le fichier IRIS.BDM.
2. Définir à l'aide de « Define Status » le statut des variables : mettre en TARGET la variable TYPE, et en INPUT les variables SEP_LENGTH, SEP_WIDTH, PET_LENGTH, PET_WIDTH.
3. Insérer le composant de discrétisation supervisée MDLPC. A ce stade, votre diagramme est le suivant :



4. Insérer de nouveau un « Define status » : mettre en TARGET la variable TYPE, et en INPUT les nouvelles variables générées par la discrétisation.



5. Il est dès lors possible d'introduire le bayesien naïf et de l'évaluer à l'aide d'une validation croisée



Le taux d'erreur associé est

Cross-validation 1

Parameters

Cross-validation parameters	
Folds	2
Trials	5

Results

CV error rate

Range	
MIN	0.0533
MAX	0.0933

Trial	Err rate
1	0.0667
2	0.0933
3	0.0600
4	0.0533
5	0.0733

Overral cross-validation error rate

Error rate	0.0693

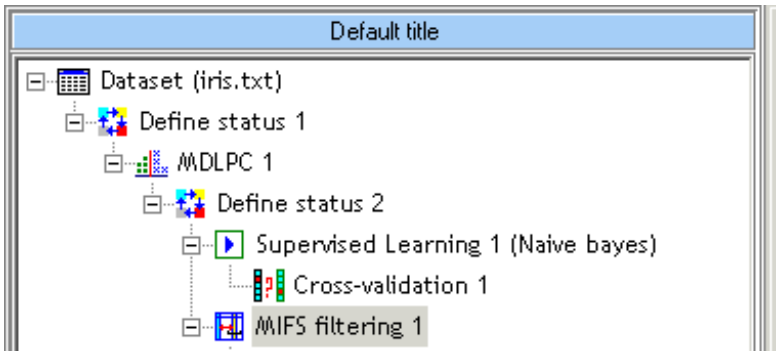
Values prediction

Value	Sensibility	Pred. error
Iris-setosa	0.9840	0.0000
Iris-versicolor	0.8840	0.0868
Iris-virginica	0.9240	0.1183

Confusion matrix

	Iris-setosa	Iris-versicolor	Iris-virginica	Sum
Iris-setosa	246	2	2	250
Iris-versicolor	0	221	29	250
Iris-virginica	0	19	231	250
Sum	246	242	262	750

6. L'idée maintenant est de déterminer s'il est possible de sélectionner un sous-ensemble des descripteurs qui permettrait d'obtenir les mêmes performances en prédiction, voire les améliorer. Nous utiliserons pour cela la méthode MIFS (Battiti et al., 1994). Nous l'insérons donc après le composant « Define Status 2 », le rôle de MIFS est de filtrer les descripteurs en sélectionnant ceux qui sont les plus pertinents pour l'apprentissage supervisé. Voici le diagramme correspondant



Les résultats montrent que seules les variables PET_LENGTH et PET_WIDTH discrétisées ont été jugées pertinentes, les autres sont exclues. Le dernier tableau indique l'ordre dans lequel les variables ont été introduites.

MIFS filtering 1

Parameters

MIFS parameters	
Beta	1.50

Results

INPUT attribute selection

INPUT selection	
Before filtering	4
After filtering	2

Keeped into INPUT selection

	Attributes
1	d_mdlpc_pet_length_1
2	d_mdlpc_pet_width_1

Removed from INPUT selection

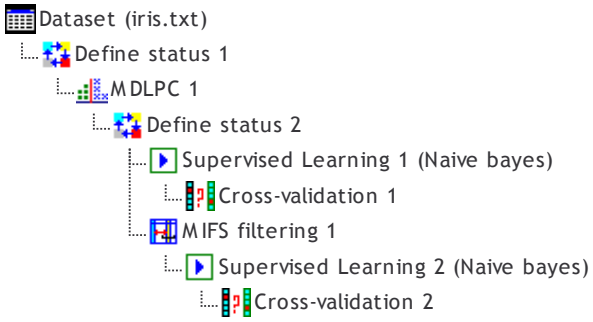
	Attributes
1	d_mdlpc_sep_length_1
2	d_mdlpc_sep_width_1

Calculations details

Selected attribute	I(Y,X/S)
d_mdlpc_pet_width_1	1.378403
d_mdlpc_pet_length_1	0.293534

Execution time : 0 ms.
Created at 18/05/2004 14:34:45

7. Pour réellement évaluer la pertinence de cette sélection, il est important de lancer de nouveau le processus d'apprentissage en incluant la sélection de variables dans la chaîne de traitements. Nous construisons donc le diagramme suivant,



Les performances du processus complet confirment la pertinence de la sélection sur ces données.

Cross-validation 2

Parameters

Cross-validation parameters	
Folds	2
Trials	5

Results

CV error rate

Range	
MIN	0.0333
MAX	0.0667
Trial	**Err rate**
1	0.0333
2	0.0400
3	0.0667
4	0.0533
5	0.0467

Overral cross-validation error rate

Error rate	0.0480

Values prediction

Value	Sensibility	Pred. error
Iris-setosa	1.0000	0.0000
Iris-versicolor	0.9200	0.0650
Iris-virginica	0.9360	0.0787

Confusion matrix

	Iris-setosa	Iris-versicolor	Iris-virginica	Sum
Iris-setosa	250	[illegible]	0	250
Iris-versicolor	0	230	20	250
Iris-virginica	0	16	234	250
Sum	250	246	254	750

Execution time : 360 ms.
Created at 18/05/2004 14:49:58