TP3 – **Sélection de modèle**

{Yann.Esposito}@lif.univ-mrs.fr

7 octobre 2005

☞ Il va être question de comparer les différentes méthodes de sélection de modèles afin d'établir un protocole général.

1 Importation des données (que du plaisir)

Allez sur le site : http://www.lif.univ-mrs.fr/~esposito/ Dans la section enseignement, téléchargez tous les fichier nécessaire au TP n°3 :

- wave-train.csv
- wave-test.csv

1. Lancez SAS ;
2. Dans le menu fichier choisissez importer ;
3. Choisir le format csv ;
4. Donnez le bon chemin de fichier ;
5. Choisissez Library : SASUSER et Member : nom du fichier sans tiret, par exemple Hugetrain.
6. Cliquez sur terminer.

Recommencez cette opération pour le fichier wave-test.csv.

2 Découpage

☞ La façon la plus immédiate de choisir un modèle est d'évaluer sa qualité directement avec sa qualité sur un ensemble de test.

2.1. Dans tout le TP les tests de qualité se feront par rapport au jeu de données test.csv. Dans un premier temps créez les nœuds `Input data source` correspondant à chacun des fichiers de données. N'oubliez de sélectionner la variable `class` et de lui attribuer le rôle de cible ainsi que le `"measurment"` interval. De plus attribuez le rôle `TEST` au nœud correpondant à l'échantillon test. Reliez le nœud de

l'ensemble d'entrainement à un nœud `Sampling` puis reliez celui-ci à un nœud `Data Parition`. Dans un premier temps mettez 100% de l'échantillon en `Train`. Reliez le nœud Data partition à un noeud de régression. Connectez le nœud correspondant à l'ensemble de test à l'ensemble de régression.

(a) Exécutez la régression linéaire en changeant à chaque étape le pourcentage du `Sampling` : 5%, 10%, 25% et 100%.
ATTENTION ne modifiez pas la valeur du `seed` tout au long du TP sauf si cela vous est demandé.
Évaluez la qualité par rapport en l'ensemble de test. Par exemple en utilisant le critère `Mean square error` obtenu en cliquant sur les résultats de la régression puis en sélectionnant l'onglet `Statistics`.

(b) Maintenant séparez chacun des échantillons d'apprentissage en deux sous-échantillon : train + validation. Puis dans la fenêtre de régression. Choisissez l'onglet Selection Method et choisissez Validation Error. Y-a-t'il eu une améliorations ?

(c) modifiez l'ordre des instances afin de modifier les différents ensembles train, validation. Faite cette opération au moins 3 fois. Constatez-vous des disparités dans les résultats ?

2.2. Décrivez les avantages et les inconvénients du découpage.

Avantages :
- facile à mettre en œuvre
- temps de calcul réduit

Inconvénients :
- on utilise pas toutes les données, en nécessite donc beaucoup :
  - deux ensembles pour l'évaluation d'un modèles
  - trois pour la sélection et l'évaluation
- sensible au découpage
- réduit les données disponibles pour le modèles, mauvaise estimation des paramètres.

3 Validation croisée

3.1. toujours sur les mêmes échantillons, utilisez de la validation croisée (onglet Selection Method de la fenêtre de régression linéaire). Évaluez la sensibilité de la validation croisée par rapport à l'ordre des éléments de l'échantillon ?

3.2. Dans SAS EM, on ne peut pas modifier le nombre de blocs de la validation croisée. Mais à votre avis, est-ce un paramètre auquel sont sensible les résultats ? Justifiez votre réponse.

Bien évidemment, l'apprentissage est sensible au nombre de blocs. Exemple, avoir un seul bloc est équivalent à utiliser le découpage.

3.3. Quel sont les conséquences d'un apprentissage par validation croisée lorsque l'on apprend avec un arbre au lieu d'apprendre avec des modèles continus comme c'est le cas pour la régression linéaire ?

On peut difficilement choisir le modèle "moyen" entre tous. Il faut alors faire des choix qui peuvent s'averer destructeurs. Par exemple, il est possible que le résultat soit aussi mauvais que celui de n'avoir fait qu'un seul des apprentissages de la validation croisée.

3.4. Donnez les avantages et les inconvénients de la validation croisée comme mode de sélection de modèle.

Avantages :
- facile à mettre en œvre
- utilise toutes les données

Inconvénients :
- sensible au découpage :
  - choix du nombre de blocs
  - choix des blocs eux-mêmes
- temps de calculs élevé
- difficile de savoir quel modèle choisir lorsqu'il s'agit d'arbre ou plus généralement de modèles discontinus.

4 Contrôle de complexité

☞ Idée de base étudier la combinaison

$$\mathcal{E} + \mathcal{C}$$

où $\mathcal{E}$ est l'erreur et $\mathcal{C}$ la complexité du modèle.
Critère de Mallows :

$$E + 2\frac{W}{N}\sigma^2$$

où
– E désigne l'erreur quadratique moyenne sur l'ensemble d'apprentissage
– W le nombre de paramètres du modèle linéaire
– N le nombre de données et
– σ^2 une estimation de la variance du bruit
Dans SAS vous avez accès à deux critères qui généralisent le critère de Mallows : AIC et SBC. Pour plus de précisions allez dans l'aide et choisissez EM Reference. Puis choisissez Regression, enfin faite une recherche sur le terme AIC.

4.1. Comparez les résultats avec ceux précédemment obtenus avec les autres méthodes de sélection de modèle.

4.2. Donnez les avantages et les inconvénient de la méthode de contrôle de complexité.

Avantages : – relativement facile à mettre en œuvre – utilise toute les données – temps de calcul additionnel négligeable Inconvénients : – AIC sélectionne des modèles trop complexes avec N grand – comportement parfois décevant quand N est 'raisonnable' – il faut estimer le bruit.

5 Conclusion

☞ Il existes d'autres méthodes de sélections de modèles. Le contrôle de complexité ; cette méthode évalue la valeur : Erreur + Complexité. Assez facile à mettre en œuvre, mais se limite généralement aux modèles linéaires. Le bootstrap ; cette méthode permet d'avoir des bornes sur l'erreur mais coûte très cher en calculs et en pratique est aussi performant que la validation croisée. Utiliser la dimension de Vapnik-Chervonenkis ; il s'agit d'une des théorie les plus avancées pour l'apprentissage automatique. En pratique pourtant il est souvent très difficile de calculer la VC-dim. De plus cette méthode done des bornes pessimistes, en grande partie car aucune hypothèse n'est faite sur les données.

Vous avez étudié les méthodes les plus utilisables en pratiques. Il s'agit maintenant de synthétiser ce que vous avez appris.

5.1. établissez un protocole général qui vous permettra de faire le bon choix parmis les différentes méthodes de sélection de modèles.

Dans le cas général, il existe d'autres façon de faire de la sélection de modèle : En pratique :
– si le temps de calcul est acceptable : validation croisée
– sinon contrôle de complexité
– sinon découpage (Train, validation, test)

✳