Travail personnel / en binômeClassification supervisée de données de biopuces

Jacques van Helden

2015-02-16

Contents

But du travail

Ce travail consiste à mener une évaluation comparative d'approches pour la classification supervisée. Le jeu de données utilisé est issu d'analyse de biopuces.

On comparera au moins deux approches qui se distingueront l'une de l'autre par un des critères suivants:

- méthode de sélection de variables (Welch's test, ANOVA, tri des variables par variance, ratio des moyennes,. . .);
- méthode de classification supervisée (LDA, QDA, SVM, KNN, . . .).

Vous mesurerez pour chacune des approches les indicateurs de performances (taux de classification correcte, taux d'erreurs, faux-positifs, faux-négatifs, . . .) en fonction du nombre de variables sélectionnées, et comparerez ces taux à ceux attendus au hasard (tests de permutation).

Vous discuterez ensuite les résultats obtenus et les forces et faiblesses des approches selectionnées.

Organisation

Le travail sera résalisé en **binômes**.

Date-limite pour la remise des rapports: lundi **19 janvier** à minuit.

Format du rapport

Comme pour le premier travail personnel. Nous attendons un rapport de ~5 pages, structuré comme un petit article scientifique (Introduction, Matériel et méthodes, Résultats et discussion, Conclusions et perspectives, Annexes éventuelles).

Ce rapport devra être remis sous deux formats.

1. Un document Rmd qui devra nous permettre d'évaluer la qualité du code et de reproduire l'analyse.
2. Un document HTML ou pdf avec le rapport généré par ce document Rmd.

Jeu de données

Il s'agit du jeu de données produit par Den Boer et al (2009) que nous avons manipulé au cours des TPs (http://pedagogix-tagc.univ-mrs.fr/courses/ASG1/). Les jeux de données sont accessibles à partir du TP "Supervised classification".

Chaque binôme choisira un sous-type de cancer de son choix, parmi les 4 qui comportent au moins 30 échantillons, et testera les performances d'une classification supervisée visant à séparer ce sous-type particulier de tous les autres.

Sélection de variables (gènes)

A priori vous pouvez choisir toute approche pertinente pour la sélection de variables:

- tri des gènes par variance;
- test de comparaison de moyenne (Student, Welch, SAM, ...) entre le sous-type d'intérêt, et les autres sous-types;
- comparaison de moyennes entre groupes multiples (ANOVA);
- tri "supervisé", en mesurant la capacité individuelle de chaque variable (gène) à discriminer les groupes, et en triant les gènes en fonction de leur pouvoir discriminant
- utilisation de ces critères soit sur les variables (gènes) initiales, soit sur les composantes principales
- ...

Note: vous pouvez éventuellement ajouter une sélection aléatoire de variables, pour comparer les courbes de hit rate avec des variables sélectionnées de façon orientée versus variables sélectionnées au hasard.

Impact du nombre de variables sélectionnées

Le travail devra comporter une évaluation de l'impact du nombre de variables sélectionnées sur le taux de classification correcte. Pour cela, vous mesurerez le taux de classification correcte en fonction du nombre de variables sélectionnées. Vous pouvez limiter cette analyse à un nombre raisonnable de nombres de variables sélectionnées, mais en les étalant pour percevoir l'impact des changements d'ordre de grandeur (par exemple: 1, 2, 5, 10, 20, 50,100, 200, 500, 1000, 2000, 5000).

Les résultats de cette analyse pourront être présentés sous forme de tableau et/ou sous forme de courbes indiquant la progression du taux de classification correcte en fonction du nombre de variables retenues.

Méthodes de classification

Durant le TP nous avons utilisé la méhode d'analyse discriminante linéaire (fonction *lda()* en R). Vous pouvez utiliser toute méthode de classification supervisée qui vous semble pertinente (y compris la LDA, si vous choisissez de comparer aumoins deux méthodes). Pour faciliter la comparaison, nous recommandons d'utiliser une méthode dont l'interface d'utilisation ressemble à celle de la *lda()*.

Evaluation des performances

Mesures de performances

L'indicateur de performance général sera le taux de classification correcte (hit rate). Si vous le désirez, vous pouvez également analyser des indicateurs particulier (taux de faux-positifs, de faux-négatifs, sensibilité, spécificité), au cas où ces indicateurs vous fourniraient des indications intéressantes pour l'interprétation des résultats

Validation croisée

Attention: toute évaluation devra se faire en mode validation croisée, en utilisant des échantillons séparés pour l'entraînement et l'évaluation. Nous vous recommandons d'utiliser des méthodes de classification qui gèrent la validation croisée de façon automatique. Par exemple, les méthodes lda() ou qda() incluent une option "CV" qui effectue automatiquement une validation croisée de type "Leave-one-out" (LOO).

Tests de permutation

Comme nous l'avons vu au TP, afin d'interpréter la qualité d'un classifieur supervisé, il est important de mesurer le tauw de classification correcte attendu au hasard. Ceci est d'autant plus important que les taux attendus au hasard peuvent s'avérer assez élevés quand les groupes d'entraînement et de test sont de tailles très différentes. Votre analyse devra donc systématiquement comparer les courbes de taux de classifications correctes obtenues avec les données réelles et les courbes obtenues avec des données permutées.

Comme pour le TP, vous pourrez tester différents modes de permutation : 1. Permutation des étiquettes (assignation aux classes d'entraînement). 2. Permutaton des valeurs d'expression au sein de chaque ligne. 3. Permutation des valeurs d'expression avec la table entière.

Nous vous demandons de choisir l'un de ces tests de permutations, et de justifier votre choix.

Précisions additionnelles

N'hésitez pas à contacter les enseignants pour obtenir des précisions supplémentaires concernant ce travail. Si des clarifications sont nécessaires, nous les ajouterons à cette page Web.