apGIS 4 : Devoir surveillé de classification supervisée

Vincent Vandewalle

30/03/2018

Vous disposez de deux heures et tous les documents sont autorisés. A la fin du devoir, votre travail est à me retourner par mail sur mon adresse Polytech'Lille, sous forme du fichier Rmd complété par vos réponses. Ce fichier pourra être complété par une copie papier.

L'objectif de ce devoir est d'étudier les facteurs qui influencent la mortalité d'un accident.

Les données sont issues de l'observatoire national interministériel de sécurité routière (extraction du 3 juillet 2012). Il s'agit de la base de données des accidents sur 6 années avec informations de géolocalisation. Le descriptif détaillé de ces données est présent dans le fichier *descriptif.pdf*.

Les données sont issue d'une version simplifiée du fichier ETALAB ACCIDENTS. Ce fichier simplifié est constitué initialement de 454.372 lignes correspondant à 454.372 accidents. Parmis ces accidents les effectifs des différentes valeurs possibles de la variable *ttue* (qui représente le nombre de morts) sont les suivants :

Valeur	0	1	2	3	4	5	7	26
Effectif	429511	23142	1465	188	48	12	5	1

Ici il y a beaucoup plus d'accidents non-mortels que d'accidents mortels (heureusement . . .).

Question 1 : Les méthodes étudiées en classification supervisées vous permettent-elles de prédire le nombre de morts (justifier) ? Pourquoi une régression linéaire classique ne serait-elle pas non plus très adaptée (justifier) ? (1,5 points)

Réponse 1 :

Pour simplifier le problème on décide de créer une nouvelle variable nommée *mort* qui prend la valeur 1 si l'accident est mortel et 0 sinon.

Question 2 : En l'état actuel la proportion d'accidents non mortels est très élevée. Quel serait le taux de bon classement obtenu si on classait tous les individus dans la classe majoritaire ? Quelle serait alors la sensibilité et la spécifité de la règle de classement ? (1,5 points)

Réponse 2 :

Enfin pour limiter la quantité de données à traiter, on tire au hasard 24.861 lignes correspondant à des accidents non mortels, et on conserve toutes les lignes correspondant à des accidents mortels. Il s'agit en fait d'une méthode d'échantillonnage appelée échantillonnage retrospectif. Son impact sur l'estimation des paramètres du modèle de régression logistique fera l'objet d'une question ultérieure.

Vous pouvez maintenant charger le fichier *accidents.Rda* resultant de ces prétraitements.

```
load("accidents.Rda")
```

Question 3 : Maintenant que vous disposez des données vous pouvez avancer davantage sur l'analyse. A l'aide d'un simple résumé des données et d'une lecture du descriptif des variables, lister les variables qui vous semblent pertinentes pour l'analyse. Donner aussi la liste des variables qu'il faut aboslument exclure de l'analyse. (2 pts)

Réponse 3 :

Question 4 : Enfin pour conclure sur les questions d'ordre général. Les données disponibles vous suffisent-elles à prédire les zones les plus dangereuses ? Pourquoi ? (1 point)

Réponse 4 :

Question 5 : Y-a-t'il une liaison significative entre la variable *lum* et la variable *mort* ? Quelle est la condition de luminosité où les accidents corporels sont les plus mortels ? (2 points)

Réponse 5 :

Question 6 : Parmis les variables explicatives *lum*, *agg*, *int*, *atm*, *col*, quelle est la plus corrélée à la variable mort ? (2 points)

Réponse 6 :

Question 7 : Ajuster un modèle de régression logistique permettant de prédire la variable *mort* en fonction de *lum*. Quelle est la modalité de la variable *lum* qui sert de modalité de référence ? Par combien est multiplié le risque d'accident mortel quand la lumière passe de "plein jour" à "crépuscule ou aube". (2 points)

Réponse 7 :

Question 8 : Ajuster un modèle de régression logistique permettant de prédire la *mort* en fonction de *lum*, *agg*, *int*, *atm* et *col*. Commenter le résultat. (1 point)

Réponse 8 :

Question 9 : Faire une sélection de variables pas à pas sur le modèle précédent. Qu'en concluez-vous ? (1 point)

Réponse 9 :

Question 10 : Evaluer les performances prédictives du modèle précédent. (2 points)

Réponse 10 :

Question 11 : Maintenant à vous de jouer, et de proposer le modèle de régression logistique le plus pertinent possible, commenter les variables retenues au final, et évaluer les performances de ce modèle. (4 points)

Réponse 11 :

Faisons maintenant un peu de mathématiques !!! Comme dit au début les proportions dans les données qui vous ont étées fournies ne sont pas représentatives des vraies proportions de mort et de non-mort. Vous allez maintenant montrer que cette modification perturbe assez peu les résultats de la régression logistique.

Question 12 En notant $f_0(x)$ et $f_1(x)$ les densités de probabilité de X dans les classes 0 et 1, puis π_0 et π_1 les proportions des classes 0 et 1, montrer que (2 points)

$$\ln \frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \ln \frac{\pi_1 f_1(x)}{\pi_0 f_0(x)}$$

Réponse 12 :

Question 13 : Supposons maintenant qu'on ait ajusté le un modèle de régression logistique mais sur des données issues d'un échantillonnage rétrospectif en imposant $\tilde{\pi}_0$ et $\tilde{\pi}_1$ les proportions des différentes classes. On notera $\tilde{P}(Y=1|X=x)$ les probabilités estimées par ce modèle. En passant par l'équation utilisée dans la question 12, quel est alors le lien entre $\ln \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$ et $\ln \frac{\tilde{P}(Y=1|X=x)}{\tilde{P}(Y=0|X=x)}$. (1 point)

Réponse 13 :

Question 14 : Enfin supposons qu'on ait ajusté un modèle de régression logistique à l'aide de l'échantillonage rétrospectif sur $\ln \frac{\tilde{P}(Y=1|X=x)}{\tilde{P}(Y=0|X=x)}$ sous la forme :

$$\ln \frac{\tilde{P}(Y=1|X=x)}{\tilde{P}(Y=0|X=x)} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_d x_d$$

alors comment peut-on en déduire le modèle de régression logistique sur $\ln \frac{P(Y=1|X=x)}{P(Y=0|X=x)}$ si les proportions π_0 et π_1 sont connues. Quels sont alors les liens entre les β_0, β_1, ..., β_d (inconnus) et les $\tilde{\beta}_0$, $\tilde{\beta}_1$, ..., $\tilde{\beta}_d$ (connus). (1 point)

Question 15 : En déduire dans le cas des données accident la correction qu'il faut appliquer aux résultats du modèle précédemment appris pour obtenir les "vrais" $P(\text{mort} = 1|X = x)$. (1 point)

Réponse 15 :

Question 16 : Ici pourquoi sur les données dont vous disposez ne pouvez-vous pas appliquer la LDA ou la QDA ? (1 point)

Réponse 16 :

Question 17 : Utiliser la fonction *naiveBayes* du package *e1071* pour apprendre le modèle prédictif. A l'aide de la documentation de la fonction dire à quelle méthode vue en cours la méthode utilisée est à relier, et expliquer l'hypothèse qui est faite ici. (1 point)