Apprentissage de données biologiques 2023-2024
Contrôle des connaissances (durée : 1h)

25 octobre 2023

Nom Prénom :

Tous les documents sont autorisés. La calculatrice est autorisée. L'usage du téléphone portable est interdit, même pour sa fonction calculatrice.

Détermination du lieu de production d'un café par spectroscopie proche infra-rouge

Un groupe industriel commercialisant du café souhaite élaborer une règle automatique de détermination du lieu de production d'un café à partir de spectroscopie proche infra-rouge. Pour cela, il souhaite s'appuyer sur un tableau de données contenant les spectres proches infra-rouge et le lieu de production, codé par un entier entre 1 et 7, de 240 échantillons de café : 50 du lieu `1`, 26 du lieu `2`, 26 du lieu `3`, 13 du lieu `4`, 22 du lieu `5`, 84 du lieu `6` et 19 du lieu `7`.

Dans la suite, Y désigne le lieu de production d'un café et $x = [x(\lambda_1), ..., x(\lambda_m)]'$ son spectre proche infra-rouge après transformation SNV (ici, $m = 1050$ et $\lambda_1 < \lambda_2 < ... < \lambda_m$ est la séquence des longueurs d'onde supports du spectre). Dans un premier temps, on choisit de construire un modèle de la probabilité qu'un café ait été produit en un lieu donné par son spectre proche infra-rouge: $\mathbb{P}_x(Y = j),\ j = 1, ..., 7$.

Question 1

Quel est le nom usuel du modèle de ce type le plus simple ? Donner l'expression mathématique du modèle et son nombre de paramètres.

Réponse

Lorsque, à l'aide de la fonction `multinom` du package `nnet` de `R`, on tente d'ajuster le modèle de la question 1, le logiciel retourne le message d'erreur suivant :

```
library(nnet)
mod <- multinom(Localisation~., data=dta)
```

```
Error in nnet.default(X, Y, w, mask = mask, size = 0, skip = TRUE, softmax = TRUE, : trop
```

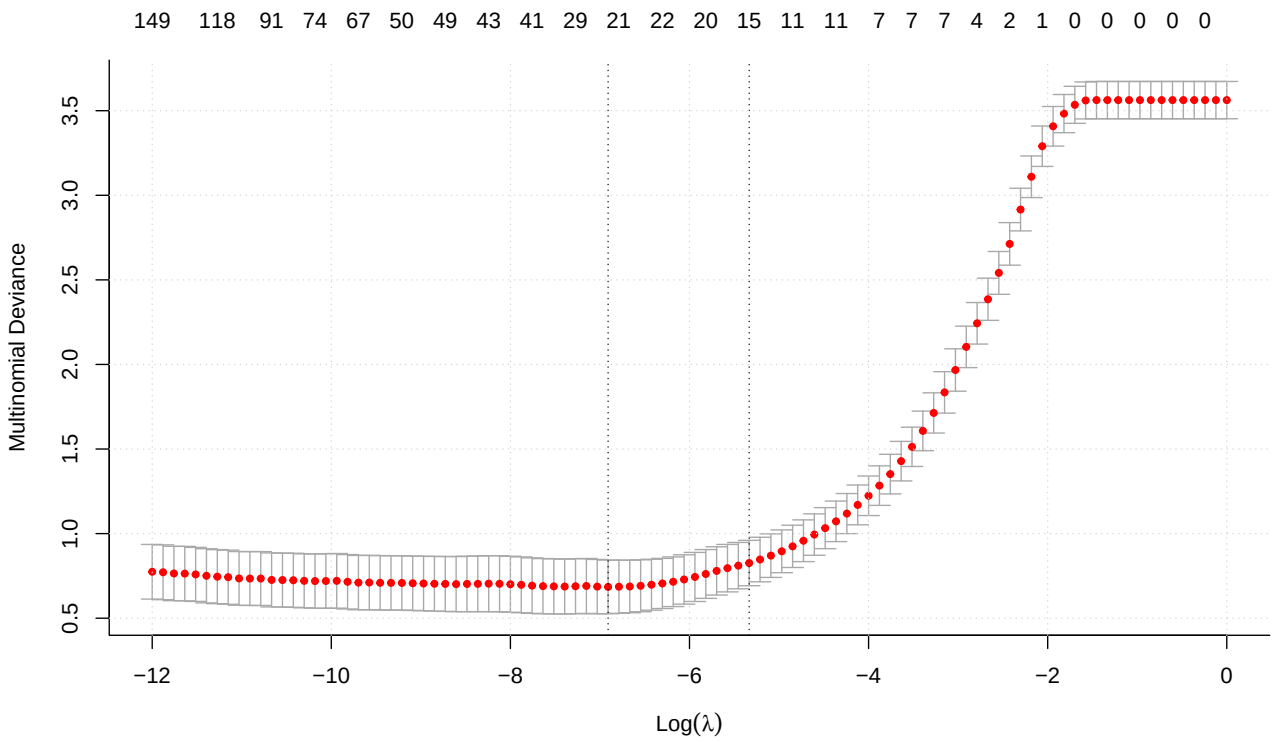
Question 2

Quel est le nom usuel du critère de qualité d'ajustement que la fonction `multinom` a tenté d'optimiser ? Expliquez pourquoi l'ajustement a échoué.

Réponse

L'estimation pénalisée permet de contourner le problème identifié à la question précédente. Dans la suite, on met en oeuvre une telle méthode d'estimation pénalisée (le terme de pénalisation est proportionnel à la somme des valeurs absolues des paramètres du modèle) pour ajuster le modèle de la question 1. Le graphique de la figure suivante montre l'évolution de la déviance résiduelle pénalisée estimée par validation croisée (10-fold), en fonction du paramètre λ permettant de contrôler le niveau de pénalisation.

```
require(glmnet)
loglambda <- seq(from=-12,to=0,length=100)
cvmod <- cv.glmnet(x=as.matrix(dta[,-1]),y=dta$Localisation,
                   type.measure="dev",alpha=1,family="multinomial",
                   lambda=exp(loglambda))
plot(cvmod,bty="l",lwd=2,col="black")
grid()
```



Question 3

D'après le graphique ci-dessus, donnez une évaluation du gain en qualité d'ajustement entre le modèle nul (sans auncune variable explicative) et le meilleur modèle (obtenu avec la valeur de λ pour laquelle la déviance résiduelle pénalisée estimée par validation croisée est la plus faible). Combien de variables explicatives y-a t'il dans ce meilleur modèle ?

Réponse

Finalement, on choisit de s'intéresser au meilleur modèle tel que défini dans la question précédente. A l'aide de ce modèle, on peut estimer les probabilités dites *a posteriori* qu'un échantillon de café ait été produit dans chacun des 7 sites, à partir de son spectre proche infra-rouge. Une règle de classification courante consiste à prédire le lieu de production de cet échantillon de café par le site de plus grande probabilité *a posteriori*.

En procédant ainsi, on commet quelques erreurs d'affectation, comme le montre la matrice de

confusion suivante, calculée à partir d'une procédure de validation croisée à 10 segments :

```
require(groupdata2)
segs <- fold(dta,k=10,cat_col="Localisation")$".folds"
   # 10-fold balanced partition of the sample
cvpred <- rep("0",times=nrow(dta))
for (k in 1:10) {
   train <- dta[segs!=k,]
   test <- dta[segs==k,]
   dta.cvlasso <- cv.glmnet(as.matrix(train[,-1]),train[,1],
                            family="multinomial",
                            type.measure="deviance",
                            lambda=exp(loglambda))
   cvpred[segs==k] <- predict(dta.cvlasso$glmnet.fit,
                              newx=as.matrix(test[,-1]),
                              type="class")[,which.min(dta.cvlasso$cvm)]
}
confusion <- table(dta$Localisation,cvpred,dnn=list("Obs.","Pred."))
confusion
```

```
    Pred.
Obs.  1  2  3  4  5  6  7
   1 48  0  1  0  0  1  0
   2  0 26  0  0  0  0  0
   3  0  0 21  0  1  4  0
   4  0  0  0 13  0  0  0
   5  0  0  1  0 19  1  1
   6  2  0  2  0  1 79  0
   7  0  0  0  1  0  0 18
```

Question 4

Donnez une estimation de l'accuracy de cette méthode de classification ?

Réponse

On espère gagner encore en performance et en stabilité de la sélection du modèle en utilisant une méthode plus adaptée pour réduire l'information portée par chaque spectre proche infrarouge. Le postulat de cette méthode est que chaque spectre $x(\lambda)$ peut être décomposé de la

façon suivante:

$$x(\lambda) \;=\; a_1 b_1(\lambda) + a_2 b_2(\lambda) + ... + a_k b_k(\lambda) + e(\lambda), \qquad (1)$$

où les fonctions $b_j(\lambda)$ sont connues, les coefficients a_j sont des paramètres inconnus et $e(\lambda)$ est une erreur d'approximation.

De manière équivalente, si $x = (x(\lambda_1), ..., x(\lambda_m))'$ désigne le vecteur des valeurs observées d'un spectre à chaque longueur d'onde, alors :

$$x \;=\; Ba + e,$$

où B est la matrice $m \times k$ dont le terme (i, j) est $b_j(\lambda_i)$, $a = (a_1, ..., a_k)'$ et $e = (e(\lambda_1), ..., e(\lambda_m))'$.

Question 5

Donnez le nom usuel d'une série de fonctions $b_j(\lambda)$ qui pourrait permettre une approximation intéressante de chaque spectre. Pour le choix de cette série de fonctions, quelle est la nature mathématique des fonctions $b_j(\lambda)$? Comment le paramètre k influence-t'il la qualité d'approximation de $x(\lambda)$?

Réponse

Question 6

Donnez l'expression du vecteur $\widehat{x} = (\widehat{x}(\lambda_1), ..., \widehat{x}(\lambda_m))'$ des valeurs ajustées de x par la méthode des moindres carrés, en fonction de B et x.

Réponse

Dans la suite, on choisit $k = 30$.

On construit donc ainsi la matrice $\hat{A}$ dont la $i$ème ligne est le vecteur $\hat{a}$ associé au spectre proche infra-rouge du $i$ème café :

```
require(splines)
basis <- bs(1:1050,df=30)
A <- matrix(0,nrow=nrow(dta),ncol=30)
for (i in 1:nrow(dta)) {
   mod <- lm(x~-1+.,data=data.frame(x=unlist(dta[i,-1]),basis))
   A[i,] <- coef(mod)
}
```

Afin de mieux comprendre comment les coefficients $\hat{a}$ peuvent permettre de caractériser les différents lieux de production de café, on choisit d'utiliser l'analyse discriminante linéaire. Le graphique de la figure suivante montre la représentation des échantillons de cafés décrits par leurs 2 premiers scores discriminants.

Le tableau suivant donne les tests de Fisher de comparaison des scores discriminants moyens des lieux de productions 3 et 6 d'une part et 4 et 5 d'autre part :

```
tests <- matrix(0,nrow=2,ncol=6)
for (j in 1:6) {
  mod <- lm(score~Localisation,data=data.frame(score=lda_scores[,j],
                                    Localisation=dta$Localisation),
```

```
            subset=dta$Localisation%in%c("3","6"))
  tests[1,j] <- anova(mod)["Localisation","F value"]
  mod <- lm(score~Localisation,data=data.frame(score=lda_scores[,j],
                                  Localisation=dta$Localisation),
            subset=dta$Localisation%in%c("4","5"))
  tests[2,j] <- anova(mod)["Localisation","F value"]
}
tests <- as.data.frame(tests)
rownames(tests) <- c("F-test 3 versus 6",
                     "F-test 4 versus 5")
kable(as.data.frame(tests),booktabs=TRUE,
      col.names = as.character(paste("LDA",1:6,sep="")))
```

	LDA1	LDA2	LDA3	LDA4	LDA5	LDA6
F-test 3 versus 6	0.114686	6.52147	29.7545	3.8827	36.7583	247.00431
F-test 4 versus 5	1.047016	1.49406	31.5975	131.5074	170.9633	7.48294

Question 7

Quel score discriminant permet le mieux de différencier les lieux de production 4 et 5 ? Même question pour les lieux de production 3 et 6 ? Quelle conséquence sur le nombre de scores discriminants à retenir dans le modèle d'analyse discriminante linéaire de Fisher ?

Réponse

La matrice de confusion associée à la règle d'affectation de la question précédente est obtenue par validation croisée (10-segments) :

```
segs <- fold(dta,k=10,cat_col="Localisation")$".folds"
   # 10-fold balanced partition of the sample
cvpred <- rep("0",times=nrow(dta))
for (k in 1:10) {
    train <- dta_spline[segs!=k,]
    test <- dta_spline[segs==k,]
    dta.lda <- lda(Localisation~.,data=train)
    cvpred[segs==k] <- predict(dta.lda,newdata=test)$class
}
```

```
confusion <- table(dta$Localisation,cvpred,dnn=list("Obs.","Pred."))
confusion
```

```
    Pred.
Obs.  1  2  3  4  5  6  7
   1 50  0  0  0  0  0  0
   2  0 26  0  0  0  0  0
   3  0  0 25  0  0  1  0
   4  0  0  0 13  0  0  0
   5  0  1  0  1 19  1  0
   6  0  0  1  1  1 81  0
   7  0  0  0  0  0  0 19
```

Pour continuer à exploiter la réduction des spectres par leurs coefficients $\widehat{a}$, on s'intéresse au modèle de régression logistique multinomiale construit à partir des coefficients $\widehat{a}$.

On propose dans un premier temps de réduire le modèle en sélectionnant le sous-ensemble des variables explicatives pour lequel le critère AIC est le plus faible.

On implémente aussi un test d'analyse de la déviance :

```
require(RcmdrMisc)
full <- multinom(Localisation~.,data=dta_spline,maxit=500,trace=-1)
select <- stepwise(full,direction="forward/backward",criterion="AIC",
                   trace=-1)
```

```
tab <- anova(select,full)[,-1]
rownames(tab) <- c("select","full")
tab
```

```
       Resid. df  Resid. Dev   Test    Df LR stat.  Pr(Chi)
select      1374 7.25099e+01
full        1254 1.20658e-04 1 vs 2   120  72.5098 0.999811
```

Question 8

Donnez une formulation aussi pratique que possible de l'hypothèse alternative du test implémenté ci-dessus ?

Réponse

Question 9

Expliquez pourquoi la colonne `Df` du tableau ci-dessus indique la valeur 120.

Réponse