Travaux encadrés de recherche

MOUDILA Marcel

MASTER 1 MATHEMATIQUES

3 juin 2022

Régression de poisson

Definition

L'équation du modèle de régression de poisson où une variable réponse est notée Y (qui suit la loi de Poisson de paramètre $\lambda > 0$), et p variables explicatives déterministes sont notées $X_1, \ldots, X_p$ est :

$$\log(\mathbb{E}(Y|X = x)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p = \log(\lambda)$$

Definition

On appelle la vraisemblance du modèle de régression de poisson, la fonction

$$\mathcal{L}(\beta, (y_i, x_i)) = \prod_{i=1}^{n} \frac{\lambda(x_i)^{y_i}}{y_i!} \exp(-\lambda(x_i))$$

l'algorithme IRLS (iteratively reweighted least squares) permet de trouver β qui maximise le logarithme de la vraisemblance ci-dessus. Muni de cet estimateur noté $\hat{\beta}$, le modèle de régression de poisson peut être employé pour faire des prédictions dans le cas où « la cible » est une variable de comptage.

avons-nous sélectionné des bonnes variables explicatives ?

L'objectif est d'évaluer l'influence de la variable explicative X_j, $1 \leq j \leq p$, sur la variable cible Y.

test de Wald

Hypothèse nulle	$H_o : \beta_j = 0$
Hypothèse alternative	$H_1 : \beta_j \neq 0$
statistique de test $\mathcal{Z}_{obs}$	$\frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$
p-value	$\mathbb{P}(\lvert\mathcal{N}(0,1)\rvert \geq \lvert\mathcal{Z}_{obs}\rvert)$

si la p-value < 0.05 pour le test de Wald concernant le coefficient β_j, alors la variable explicative X_j influence significativement la variable réponse Y, on décide alors de supprimer cette variable et on refait la mise en oeuvre de la régression de poisson sans cette variable.

est-ce que prédire via la régression de poisson serait pertinent ?

résidus de déviance

Hypothèse nulle	$H_0 : \lambda(x) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$
Hypothèse alternative	$H_1 : \lambda(x) \neq \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)$
déviance résiduelle	$\hat{D}_i = \pm\sqrt{2\left(Y_i \log\left(\frac{Y_i}{\hat{\lambda}(x_i)}\right) - (Y_i - \hat{\lambda}(x_i))\right)}$
χ^2_{obs}	$\sum_{i=1}^{n} \hat{D}_i^2$
p-value :	$\mathbb{P}(\text{chi-deux}(n-(p+1)) \geq \chi^2_{obs})$

Si la p-value > 0.05, le modèle est pertinent par rapport aux données, c'est-à-dire H_0 est acceptée, et donc qu'il y'a une fonction de lien linéaire (ici la fonction de lien est "log") entre la variable cible et les variables explicatives, car on retrouve l'équation du modèle de régression de poisson définit au premier transparent.

est-ce que certaines valeurs de la variable cible dans les données seraient anormales ?

la détection des valeurs aberrantes dans les données est cruciale car ces valeurs peuvent avoir une influence négative dans les estimations des paramètres de notre modèle de régression. Il faut donc détecter ces observations.

Definition

Pour tout $i = 1, \ldots, n$, on appelle le résidu de Pearson de l'observation i l'estimateur donné par

$$\hat{\epsilon}_i = \frac{Y_i - \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi})}{\sqrt{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi})}} = \frac{Y_i - \hat{\lambda}(x_i)}{\sqrt{\hat{\lambda}(x_i)}}$$

On note $W = \text{diag}(\hat{\lambda}(x_1), \ldots, \hat{\lambda}(x_n)$

est-ce que certaines valeurs de la variable cible dans les données seraient anormales ?

Definition

On appelle le résidu standardisé de Pearson de l'observation i la réalisation e_i^* de

$$\hat{\epsilon_i^*} = \frac{\hat{\epsilon_i}}{\sqrt{1 - [W^{\frac{1}{2}} X (X^t W X)^{-1} X^t W^{\frac{1}{2}}]_{i,i}}}$$

Si $|e_i^*| > 2$, alors on envisage que l'observation i a des valeurs aberrantes.

Mais il vaut mieux ne pas se hâter à supprimer cette observation , éventuellement se renseigner si une erreur n'avait pas été faite dans le jeux des données.

tout est ok ? place à la prédiction...

pour prédire $\hat{y_{n+1}}$ d'une nouvelle observation (si notre modèle validé contenait déjà n observations), on estime $\hat{\lambda}(x_{n+1})$ par

$$\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1,n+1} + \ldots + \hat{\beta}_p x_{p,n+1})$$

Definition

Donc

$$p_k = \mathbb{P}(Y = k) = \exp(-a)\frac{a^k}{k!}$$

où on a posé

$$k = \hat{y_{n+1}} \in \mathbb{N}$$

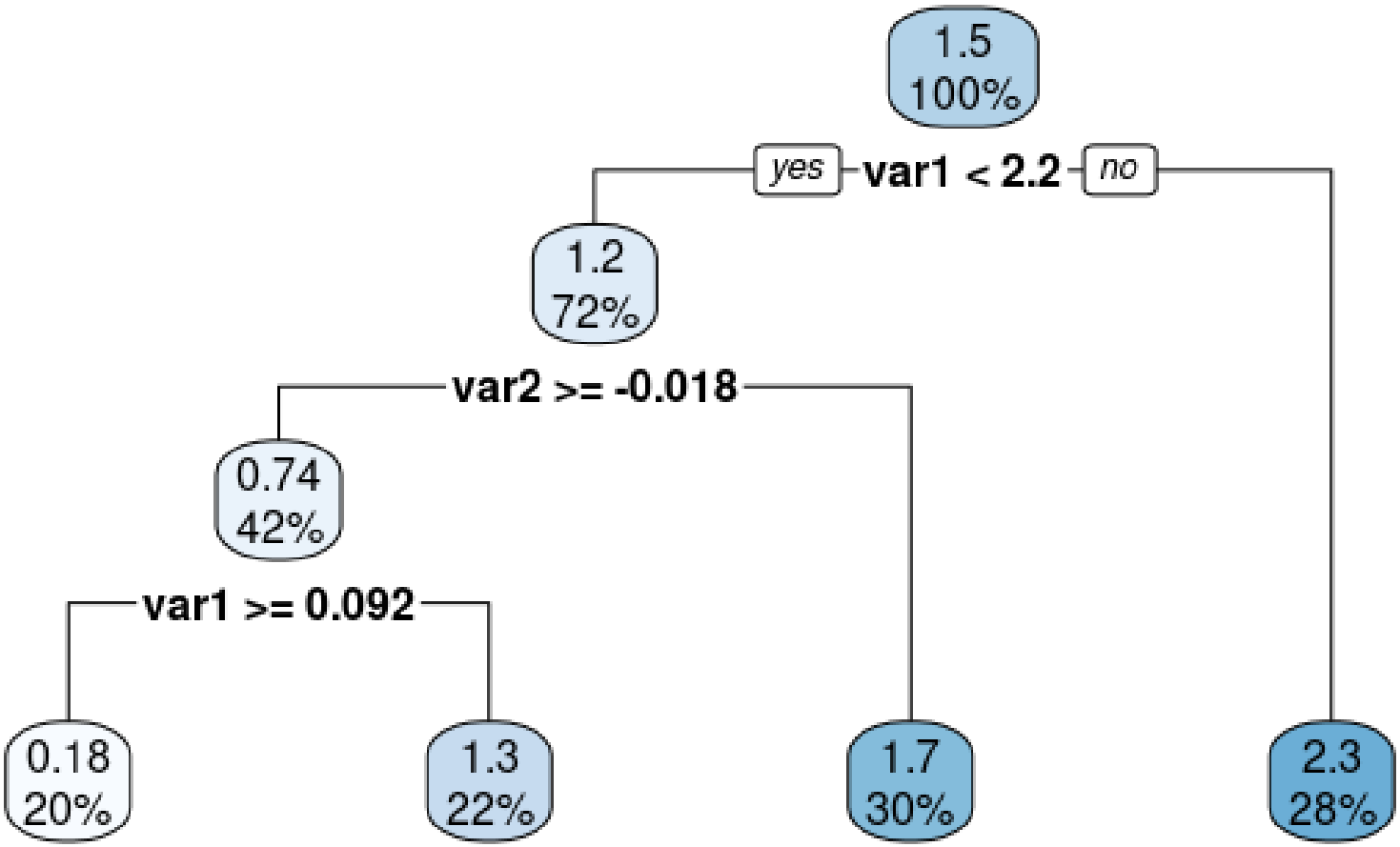
et

$$a = \hat{\lambda}(x_{n+1})$$

traduction

avec probabilité p_k, $\hat{y_{n+1}}$ prend la valeur k

exemple d'arbre de décision



Les forêts aléatoires

une forêt aléatoire est constituée en pratique d'au moins 500 arbres de décision.

Buts :

- prédire une valeur de la variable « cible » Y, qui peut être soit discrète (comme en régression de poisson) , soit continue
- prédire une modalité de la variable « cible » Y , lorsque Y est qualitative

construction

- On sélectionne au hasard et **avec remise** m couples d'observations parmi $(y_1, x_1), \ldots, (y_n, x_n)$ où $x_i = (x_{i1}, \ldots, x_{ip})$, et $m \simeq \frac{2n}{3}$
- On sélectionne **sans remise** q variables parmi $X_1, \ldots, X_p$, où $q \simeq \sqrt{p}$
- On construit 1 arbre de décision avec ces sélections.
- On répète le procédé N fois, où en pratique N vaut 500

tout est ok ? place à la prédiction...

Y quantitative

Étant donné une forêt aléatoire de régression, une valeur prédite de y_{n+1} pour un individu dont on connaît les réalisations de $X_1, \ldots, X_p$ est obtenue en faisant la moyenne des valeurs prédites de y_{n+1} par chaque arbre de la forêt.

Y qualitative

Étant donné une forêt aléatoire de classification, une modalité plausible de y_{n+1} est donnée par un vote majoritaire parmi les modalités plausibles de y_{n+1} obtenues par chaque arbre de la forêt.

comment construit-on un arbre de décision ?

Plusieurs ARTICLES de recherche dont :

- RAKOTOMALALA, Ricco. Arbres de décision. Revue Modulad, 2005, vol. 33, p. 163-187.

Plusieurs MÉTHODES dont :

- CHAID (CHi-squared Automatic Interaction Detector),Gordon V. Kass. (1980)
- CART (Classification And Regression Tree),Breiman et col. (1984)
- ID3 (Iterative Dichotomiser 3)
- C4.5, et C5 (successeurs d'ID3),Quinlan. (1993)

Jeux des données UEFA Euro Cup : kaggle

43 Results Sort by: Relevancy

Dataset

UEFA Euro Cup 2020

by Marco Carujo

a year ago • 23 kB • 29

UEFA Euro Cup 2020

Dataset

UEFA Euro Cup (1960-2016)

by Jay

a year ago • 8 kB • 25

UEFA Euro Cup (1960-2016)

Dataset

UEFA Euro Championship

by Mohammad Essam

2 years ago • 94 kB • 16

UEFA Euro Championship

Un projet prochain serait d'appliquer la régression de poisson et les forêts aléatoires sur les jeux des données des coupes d'EURO 2014 à 2016 pour prédire le vainqueur de la coupe d'EURO 2020 (le vainqueur était l'Italy)

Je vous remercie ! ! !