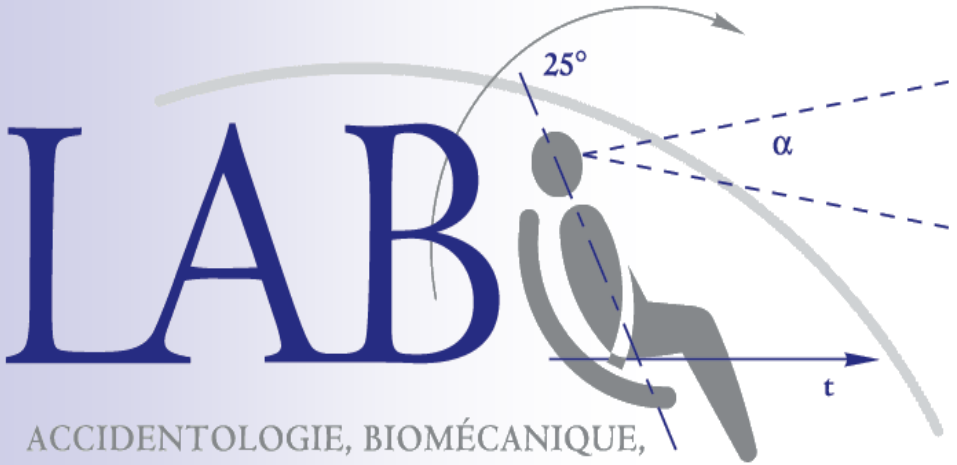
Indice de confiance non paramétrique pour la prédiction et le classement en sécurité routière

S. Doucet [1,2]

avec A. Chambaz [2], C. Chauvel [1]

[1] Laboratoire d'accidentologie et de biomécanique (LAB)

[2] Modal'X, Université Paris-Ouest Nanterre



ACCIDENTOLOGIE, BIOMÉCANIQUE, COMPORTEMENT HUMAIN

05 Avril 2016

Introduction : classement global en sécurité routière

Problématique de la thèse : généraliser et enrichir le classement

Définition et inférence d'une mesure d'importance de variables

Conclusion et perspectives

Introduction : classement global en sécurité routière

Problématique de la thèse : généraliser et enrichir le classement

Définition et inférence d'une mesure d'importance de variables

Conclusion et perspectives

Introduction 1/4 : contextes et objectifs

- **Contexte :** volonté des constructeurs automobiles d'obtenir
  - un **positionnement** d'un nouveau modèle de véhicule
  - en termes de protection offerte

    en accidentologie réelle
  - le plus rapidement possible après la sortie sur le marché
- **Objectif :**
  - en utilisant les données d'accidentologie réelles
    1. classement **contextuel** en *sécurité secondaire* des modèles automobiles
       (cf : https://hal.archives-ouvertes.fr/hal-01194515 bientôt accepté à JRRS-C)
    2. classement **global** en *sécurité offerte* des modèles automobiles
       (cf. https://hal.archives-ouvertes.fr/hal-01359225)
  - On va donc exploiter des **relevés d'accidents** sur les routes françaises

Introduction 2/4 : les données BAAC*

- Pour obtenir BAAC* :
  - données nationales d'accidents corporels BAAC (Bulletin d'Analyse des Accident Corporels)
  - restriction à des accidents avec un VP (véhicule particulier) seul en cause ou des accidents à 2 VPs
  - variable binaire pour la gravité :
    - 1 : pour occupant Tué ou BG (hospitalisé plus de 24 heures)
    - 0 : pour occupant BL (hospitalisé moins de 24h) ou Indemne
  - adossement à des données de **parc** (descriptif de classes de véhicules)
  - création par le LAB de la base de données **BAAC*** (décrit les circonstances d'accident y compris le véhicule)

Introduction 3/4 : modélisation des données BAAC*

- Observations $O^1, \ldots, O^n$ indépendantes et identiquement distribuées de loi P représentent n accidents
  - $O^i = (W^i, X^i, Y^i) \sim P$
  - $W^i \in \mathcal{W}$ représente les données du contexte, le profil de l'impliqué et celui du conducteur
  - $X^i \in \mathcal{X}$ représente la "classes générationnelles" CG du modèle de véhicule
  - $Y^i \in \{0, 1\}$ est le degré de gravité de l'impliqué

Introduction 4/4 : obtention d'un classement

- On cherche à **classer** les différentes CG par degré de sécurité offerte
- On utilise un **score** de sécurité calculé par apprentissage
- La CG x d'un véhicule est définie par son :
  1. segment (voiture familiale, minivan, etc.)
  2. année de mise en circulation
  3. année de conception
  4. 4 variables numériques et catégorielles
- Pour chaque CG x et chaque contexte w on détermine un **score contextuel** $s(x, w)$: la CG x est plus sûre que x' dans le contexte w si $s(x, w) < s(x', w)$.
- Pour toute CG x, on détermine un **score global** $\widetilde{s}(x)$ en agrégeant les scores contextuels $s(x, w)$ de x.
- Une CG x est dite plus sûre que x' "globalement" si $\widetilde{s}(x) < \widetilde{s}(x')$.

Introduction : classement global en sécurité routière

Problématique de la thèse : généraliser et enrichir le classement

Définition et inférence d'une mesure d'importance de variables

Conclusion et perspectives

Problématique de la thèse 1/2 : tâches principales

1. Définition et inférence de l'**importance** des variable
2. Meilleure exploitation des **bases de données**
   - correction du sous-enregistrement du BAAC
   - exploitation d'autres bases de données (VOIESUR, EDA, etc.)
3. Détermination d'un **indice de confiance** pour les classements
4. Optimisation algorithmique
   - par un meilleur choix d'algorithmes (apprentissage "en ligne")
   - par une optimisation du code R

Problèmatiques de la thèse 2/2 : pistes de résolution

- Pour l'instant :
  1. des pistes sérieuses sont envisagées pour 1. et 2..
  2. 1 sera présenté ci-dessous
  3. pour 2., on envisage de faire de l'**imputation**
  4. pour 3., les pistes n'ont pas encore été explorées
  5. pour 4., on formule le problème de manière générale en s'inspirant d'une présententation de L. Bottou, F. E. Curtis, J. Nocedal (https://arxiv.org/abs/1606.04838) et d'un article de T. Chen et C. Guestrin (https://arxiv.org/abs/1603.02754). On envisage aussi de faire de l'apprentissage **en ligne** et d'utiliser la plateforme de calcul H2O.
- Nous nous concentrons dans la suite sur le problème de l'importance des variables

Introduction : classement global en sécurité routière

Problématique de la thèse : généraliser et enrichir le classement

Définition et inférence d'une mesure d'importance de variables

Conclusion et perspectives

Définition d'une mesure d'importance (1/3)

- Inspiré de la biostatistique (cf. D. Benkeser, S. D Lendle, C. Ju, M. Van Der Laan, rapport technique à paraître)
- Principe :
  - l'importance d'une variable est liée à la **réduction du risque** qu'elle engendre
  - une variable significative réduit le risque du classement
  - on s'inspire du coefficient R^2 de la régression linéaire :

$$R^2 = 1 - \frac{\sum_{j=1}^{n}(y_j - \widehat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \overline{y}_j)^2}$$

qui compare le risque du modèle choisi avec le risque de l'algorithme ψ par moyenne empirique

Definition d'un mesure d'importance (2/3)

- On définit une mesure du risque relatif de l'algorithme ψ par :

$$R_0^2(\psi) = 1 - \frac{E_0[(Y - \widehat{\psi}(X))^2]}{E_0[(Y - E_0[Y])^2]}$$

- On va
  1. estimer $R_0^2(\psi)$ pour différents choix de variables
  2. comparer ces différents valeurs et en déduire les variables les plus importantes
- Plus une variabl est importante plus son exclusion des données fait **décroître** $R_0^2(\psi)$

Définition d'une mesure d'importance 3/3

- On définit l'importance de la covariable X_k dans $X = (X_1, \ldots, X_r)$
- On note $X_{-k} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_r)$ les covariables **sans** X_k
- On introduit $\widehat{\psi}_{-k}$ l'estimateur entraîné sur X_{-k} seulement
- On calcule :

$$R_0^2(\psi) = 1 - \frac{E_0[(Y - \widehat{\psi}(X))^2]}{E_0[(Y - E_0[Y])^2]},$$

$$R_0^2(\psi_{-k}) = 1 - \frac{E_0[(Y - \widehat{\psi}_{-k}(X_{-k}))^2]}{E_0[(Y - E_0[Y])^2]}$$

- L'importance de X_k dans $X = (X_1, \ldots, X_r)$ est donnée par :

$$\Delta_{-k} = R_0^2(\psi) - R_0^2(\psi_{-k})$$

Inférence de la mesure d'importance (1/3)

- Pour connaître ces quantités, il faut connaître $\mathbb{P}_0$: R_0^2 et Δ sont **inconnus**, on va donc les estimer, par **validation croisée**
- En validation croisée, on sépare les données en
  1. *échantillon d'apprentissage* et
  2. *échantillon de validation*
- On évite ainsi le sur-apprentissage et la sous-estimation du risque
- La validation croisée est coûteuse en calculs (on repasse plusieurs fois sur le données)
- D'où l'importance de l'optimisation du calcul (problématique 5.)

Inférence de la mesure d'importance (2/3)

- On obtient par validation croisée, les estimateurs respectifs $\mathcal{E}$ et $\mathcal{E}_{-k}$ de $E_0[(Y - \widehat{\psi}(X))^2]$ et $E_0[(Y - \widehat{\psi}_{-k}(X_{-k}))^2]$
- On en déduit un intervalle de confiance de niveau $1 - \alpha$ de la forme

$$\left[1 - \exp\left(\log\left(\frac{\mathcal{E}}{\mathcal{E}_{-k}}\right) \pm z_{1-\alpha/2}\frac{\sigma_n^2}{n^{1/2}}\right)\right]$$

où σ_n^2 est un estimateur consistant de variance de $\log\left(\frac{\mathcal{E}}{\mathcal{E}_{-k}}\right)$ et $z_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale.

Utilisation de la mesure d'importance

- On rappelle qu'on cherche à **classer** des Classes Générationelles (CG) de véhicules par sécurité offerte
- On **définit** une CG par 7 variables
- On **applique** la démarche de mesure d'importance de variable aux variables qui définissent la CG
- On en déduit une **description "optimale"** d'une classe de véhicules en ne retenant que les variables les plus importantes, en nombre suffisant pour décrire efficacement la CG

Introduction : classement global en sécurité routière

Problématique de la thèse : généraliser et enrichir le classement

Définition et inférence d'une mesure d'importance de variables

Conclusion et perspectives

Conclusion

- Conclusion
  - on obtient un score de sécurité, contextuel puis non contextuel, pour des CG de véhicules
  - on classe les CG par score
  - on veut améliorer et enrichir le score de différentes manières
  - notamment en déterminant la définition optimale d'une CG d'après les données
- Perspectives
  - préciser et généraliser le score en exploitant mieux les bases de données
  - contrôler les scores par un indice de confiance (non paramétrique)
  - optimiser le calcul effectif du score
  - compléter le classement par de la prédiction

Conclusion

- Conclusion
  - on obtient un score de sécurité, contextuel puis non contextuel, pour des CG de véhicules
  - on classe les CG par score
  - on veut améliorer et enrichir le score de différentes manières
  - notamment en déterminant la définition optimale d'une CG d'après les données
- Perspectives
  - préciser et généraliser le score en exploitant mieux les bases de données
  - contrôler les scores par un indice de confiance (non paramétrique)
  - optimiser le calcul effectif du score
  - compléter le classement par de la prédiction

Merci de votre attention !

Références

1. Zaïd Ouni, Christophe Denis, Cyril Chauvel, Antoine Chambaz. Contextual ranking by passive safety of generational classes of light vehicles. 2016. https://hal.archives-ouvertes.fr/hal-01194515
2. Zaïd Ouni, Cyril Chauvel, Antoine Chambaz. From contextual to global rankings by passive safety of generational classes of light vehicles. 2016. https://hal.archives-ouvertes.fr/hal-01359225
3. Benkeser, David ; Lendle, Samuel D. ; Ju, Cheng ; and van der Laan, Mark J., "Online Cross-Validation-Based Ensemble Learning" (October 2016). U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 355. http://biostats.bepress.com/ucbbiostat/paper355
4. Bottou, Léon ; Curtis, Frank, E ; and Nocedal, Jorge, "Optimisation Methods for Large-Scale Machine Learning", arXiv :1606.04838v1 [stat.ML] https://arxiv.org/abs/1606.04838
5. Tianqi, Chen ; and Guestrin, Carlos, "XGBoost : A scalable tree-boosting system", arXiv : 1603.02754v1 [cs.LG] https://arxiv.org/abs/1603.02754