Sélection Bayésienne de Modèles avec Prior Dépendant des Données

Marc Boullé *

* Orange Labs, 2 avenue Pierre Marzin, 22300 Lannion
marc.boulle@orange.com,
http://perso.rd.francetelecom.fr/boulle/

Résumé

Cet article analyse la consistance asymptotique des modèles en grille appliqués à l'estimation de densité jointe de deux variables catégorielles. Les modèles en grille considèrent un partitionnement des valeurs de chacune des variables, le produit Cartésien des partitions formant une grille dont les cellules permettent de résumer la table de contingence des deux variables. Le meilleur modèle de co-partitionnement est recherché au moyen d'une approche MAP (maximum a posteriori), présentant la particularité peu orthodoxe d'exploiter une famille de modèles et une distribution a priori de ces modèles qui dépendent des données. Ces modèles sont par nature des modèles de l'échantillon d'apprentissage, et non de la distribution sous-jacente. Nous démontrons la consistance de l'approche, qui se comporte comme un estimateur universel de densité jointe convergeant asymptotiquement vers la vraie distribution jointe.

Summary

This paper studies the asymptotic consistency of the data grid models applied to the joint density estimation of two categorical variables. The data grid models consider the grouping of the values of each variable. The Cartesian product of these partitions forms a grid whose cells provide a summary of the contingency table of the two variables. The best bivariate grouping model is searched by the mean of a MAP (maximum a posteriori) approach, with the heretic property of exploiting both a model family and a prior distribution that are data dependent. These models are in essence models of the data sample, not of the underlying distribution. We demonstrate the consistency of the approach, which behaves as a universal estimator of joint density that asymptotically converges towards the true joint distribution.