Les modèles de mélange, un outil utile pour la classification semi-supervisée

Vincent Vandewalle[1,2]

[1] Laboratoire Paul Painlevé UMR CNRS 8524
Université Lille I
59655 Villeneuve d'Ascq Cedex, France
vincent.vandewalle@math.univ-lille.fr
[2] INRIA

Résumé En classification supervisée, la règle de classement est apprise à partir d'un échantillon d'apprentissage généralement constitué de données classées. Dans la plupart des cas l'obtention de la classe est plus coûteuse que l'obtention de covariables associées à la classe d'où l'intérêt d'apprendre une règle de prédiction de la classe à partir de ces covariables. Ainsi dans de nombreuses situations beaucoup de données non classées, obtenues à un coût relativement faible, sont disponibles en plus des données classées. Au cours des dernières années la classification semi-supervisée, qui fait usage des données non classées pour améliorer la précision de la règle de classement apprise, a connu un essor important, ceci notamment dans la communauté du Machine Learning. Les modèles génératifs, qui modélisent la distribution jointe de la classe et des covariables, permettent de prendre naturellement en compte l'information apportée par les données non classées dans l'apprentissage de la règle de classement. Dans cet article nous dressons un panorama de la classification semi-supervisée et nous détaillons sa mise en oeuvre dans le cadre des modèles génératifs.
Mots-clés : données manquantes, modèles de mélange, algorithme EM, analyse discriminante, validation croisée.

Abstract In supervised classification, the classification rule is learnt from a learning sample generally composed of labeled data. In most settings obtaining the label is more expensive than obtaining covariates linked with the label, hence the interest to learn a prediction rule of the label given these covariates. So, in many settings a lot of unlabeled data, obtained at a relatively low cost, are available in addition to labeled data. Over past years the semi-supervised classification, which uses unlabeled data in order to improve the classification rule accuracy, has known a great development, especially in Machine Learning community. Generative models, which model the joint distribution of the label and of the covariates, allow to naturally take into account information contained in unlabeled data when learning the parameters of the model. In this article we give a survey of semi-supervised classification and we detail how to use it with generative models.
Keywords: missing data, mixture models, EM algorithm, discriminant analysis, cross-validation.