Le Graphe Génératif Gaussien

Pierre Gaillard[1], Michaël Aupetit[2], Gérard Govaert[3]

[1] CEA, DAM, DIF, F-91297 Arpajon, France
[2] CEA, LIST, F-91191 Gif-sur-Yvette cedex, France
[3] UTC, U.M.R. C.N.R.S. 6599 Heudiasyc, F-60205 Compiègne Cedex, France

Résumé Un nuage de points est plus qu'un ensemble de points isolés. La distribution des points peut être gouvernée par une structure topologique cachée, et du point de vue de la fouille de données, modéliser et extraire cette structure est au moins aussi important que d'estimer la seule densité de probabilité du nuage. Dans cet article, nous proposons un modèle génératif basé sur le graphe de Delaunay d'un ensemble de prototypes représentant le nuage de points, et supposant un bruit gaussien. Nous dérivons un algorithme pour la maximisation de la vraisemblance des paramètres, et nous utilisons le critère BIC pour sélectionner la complexité du modèle. Ce travail a pour objectif de poser les premières pierres d'un cadre théorique basé sur les modèles génératifs statistiques, permettant la construction automatique de modèles topologiques d'un nuage de points.

Keywords : connexité, graphe de Delaunay, modèle génératif, algorithme EM, critère BIC

1 Introduction

Dans les problèmes d'apprentissage statistique, on suppose que les données sont générées par une densité de probabilité $P : \mathbb{R}^D \rightarrow \mathbb{R}^+$. Cependant, le processus sous-jacent de génération des données défini par la fonction P, possède dans de nombreux cas d'intérêt moins de degrés de liberté que l'espace d'observation de dimension D. La formalisation de cette intuition est de supposer que les données sont sûres ou proches d'un ensemble de variétés, appelées *variétés principales* [1], chacune ayant une dimension intrinsèque inférieure à la dimension de l'espace d'observation.

Etant donné un ensemble $\underline{x}$ de M points observés, dans un espace euclidien à D dimensions, les méthodes statistiques permettent de résoudre des problèmes très généraux de discrimination, classification ou régression, en estimant la densité de probabilité de cet ensemble (modèles de mélange [2], méthodes à noyau [3]). Bien que la fonction densité de probabilité contienne la totalité de l'information extractible de la population dont le nuage de points est un échantillon, celle-ci ne rend pas explicite l'information géométrique et topologique relatives aux variétés principales. Pourtant, si l'on suppose qu'une structure existe dans les données, l'extraire et la caractériser à partir de la densité sont aussi importants que d'estimer la densité de probabilité elle-même. Par exemple, dans le contexte d'un problème de classification, la connexité de cette structure semble être le moyen naturel pour définir des groupes homogènes. L'intérêt d'utiliser cette structure sous-jacente qui gouverne la distribution des données est majeur, puisque celle-ci peut être aussi utilisée pour analyser [4], visualiser [5], discriminer les données [6].

De manière générale, on pourrait extraire des caractéristiques *géométriques* de cette structure telles que la position relative de ses différentes parties, mais aussi des caractéristiques dites *topologiques* telles que la dimension intrinsèque ou la connexité.