Apprentissage incrémental anytime d'un classifieur Bayésien naïf pondéré

Carine Hue, Marc Boullé, Vincent Lemaire

Orange Labs Lannion, 2 avenue Pierre Marzin, 22300 Lannion

Résumé. Nous considérons le problème de classification supervisée pour des flux de données présentant éventuellement un très grand nombre de variables explicatives. Le classifieur Bayésien naïf se révèle alors simple à calculer et relativement performant tant que l'hypothèse restrictive d'indépendance des variables conditionnellement à la classe est respectée. La sélection de variables et le moyennage de modèles sont deux voies connues d'amélioration qui reviennent à déployer un prédicteur Bayésien naïf intégrant une pondération des variables explicatives. Dans cet article, nous nous intéressons à l'estimation directe d'un tel modèle Bayésien naïf pondéré. Nous proposons une régularisation parcimonieuse de la log-vraisemblance du modèle prenant en compte l'informativité de chaque variable. La log-vraisemblance régularisée obtenue étant non convexe, nous proposons un algorithme de gradient en ligne qui post-optimise la solution obtenue afin de déjouer les minima locaux. Les expérimentations menées s'intéressent d'une part à la qualité de l'optimisation obtenue et d'autre part aux performances du classifieur en fonction du paramétrage de la régularisation.

1 Introduction

Du fait de l'accroissement continu des capacités de stockage, la capture et le traitement des données ont profondément évolué durant ces dernières décennies. Il est désormais courant de traiter des données comprenant un très grand nombre de variables et les volumes considérés sont tels qu'il n'est plus forcément envisageable de pouvoir les charger intégralement : on se tourne alors vers leur traitement en ligne durant lequel on ne voit les données qu'une seule fois. Dans ce contexte, on considère le problème de classification supervisée où Y est une variable catégorielle à prédire prenant J modalités $C_1, \ldots, C_J$ et $X = (X_1, \ldots, X_K)$ l'ensemble des K variables explicatives, numériques ou catégorielles. On s'intéresse à la famille des prédicteurs de type Bayésien naïf. L'hypothèse d'indépendance des variables explicatives conditionnellement à la variable cible rend les modèles directement calculables à partir des estimations conditionnelles univariées de chaque variable explicative. Pour une instance n, la probabilité de prédire la classe cible C conditionnellement aux valeurs des variables explicatives se calcule alors selon la formule :

$$P_w(Y = C|X = x^n) = \frac{P(Y = C)\prod_{k=1}^{K} p(x_k^n|C)^{w_k}}{\sum_{j=1}^{J} P(C_j)\prod_{k=1}^{K} p(x_k^n|C_j)^{w_k}} \quad (1)$$