Projet 1

Jean-Thomas Baillargeon
Christopher Blier-Wong
Pour le cours STT-7330
Méthode d'analyse des données

Présenté le 25 mars 2018 à la professeure

Anne-Sophie Charest

Département de mathématiques et de statistiques
Faculté des sciences et de génie
Université Laval

Jean-Thomas Baillargeon
Christopher Blier-Wong
Faculté des sciences et de génie
École d'actuariat
Université Laval
Hiver 2018

Table des matières

1 Introduction

L'objectif de l'analyse par composantes principales est d'obtenir une représentation des données dans un espace plus restreint en conservant la plus grande quantité d'information possible. Plus précisément, on considère les combinaisons linéaires des variables mesurées pour l'espace restreint. Par contre, on observe souvent que la relation entre les données n'est pas linéaire. L'analyse par composantes par noyau généralise ce modèle pour des combinaisons non-linéaires des attributs. Au lieu de faire une décomposition par valeurs et vecteurs propres sur la matrice de covariance des données centrées

$$\Sigma = Var(X) = \frac{1}{p}\sum_{j=1}^{p} \mathbf{x}_j \mathbf{x}_j^T,$$

on fait une décomposition par valeurs et vecteurs propres sur la matrice de covariance des données projetées sur un nouvel espace d'attributs.

2 Fonction noyau

L'idée de l'espace des attributs est de projeter les données originales par une fonction non linéaire vers un nouvel espace. Formellement, on a

$$\begin{aligned} \Phi : \mathbb{R}^N &\to \mathcal{F} \\ \mathbf{x} &\mapsto \Phi(\mathbf{x}) \end{aligned}$$

où les données $\mathbf{x}_1, \mathbf{x}_2 \ldots, \mathbf{x}_n \in \mathbb{R}^N$ est projeté vers un espace d'attributs $\mathcal{F}$ [Muller et al., 2001]. Souvent, la dimension de $\mathcal{F}$ est beaucoup plus élevée que l'espace originale. L'apprentissage statistique peut maintenant être fait sur les données $(\Phi(\mathbf{x}_1), y_1), (\Phi(\mathbf{x}_2), y_2), \ldots, (\Phi(\mathbf{x}_n), y_n)$.

Le produit scalaire entre deux espaces d'attributs peut être reformulé en terme d'une fonction noyau k par

$$\mathrm{k}(\mathbf{x}, \mathbf{y}) = (\Phi(\mathbf{x}) \cdot \Phi(\mathbf{y})).$$

Dans plusieurs problèmes d'apprentissage, le "truc du noyau" permet d'éviter de calculer directement les nouvelles données $\Phi(\mathbf{x}_n)$. En effet, pour certains algorithmes d'apprentissage,
car on peut reformuler les équations de mise à jour par le produit scalaire entre différentes données et ainsi les remplacer par la fonction de noyau. Des exemples de noyaux communément utilisés sont présentés dans la table 1.

Nom	$k(\mathbf{x}, \mathbf{y})$
Gaussien (RBF)	$\exp\left(\frac{-\|\|\mathbf{x}-\mathbf{y}\|\|^2}{c}\right)$
Polynomial	$((\mathbf{x} \cdot \mathbf{y} + \theta))^d$
Sigmodoidal	$\tanh(\kappa(\mathbf{x} \cdot \mathbf{y}) + \theta)$
Multiquadrique inversé	$\frac{1}{\sqrt{\|\|\mathbf{x}-\mathbf{y}\|\|^2+c^2}}$

TABLE 1 – Noyaux communs

3 PCA avec noyau

4 Application pratique

5 Conclusion

Références

[Muller et al., 2001] Muller, K.-R., Mika, S., Ratsch, G., Tsuda, K., and Scholkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE transactions on neural networks*, 12(2) :181–201.