Componentes de Influencia y sus aplicaciones a la clasificación

Andrés Jiménez Jiménez[1], *José María Gutiérrez Pérez*[2], *Francisco Alvarez Gonzalez*[3]

[1]andres.jimenez@uca.es, CITI., Univ. Cádiz
[2]josema.gutierrez@uca.es, Dpto. de Estadística e IO, Univ. Cádiz
[3]francisco.alvarez@uca.es, Dpto. de Estadística e IO, Univ. Cádiz

Resumen

El presente estudio plantea una generalización de la matriz de covarianza asociada a muestras supervisadas que denominamos Matriz de Dispersión Inducida. La descomposición espectral de dicha matriz es local a cada punto del espacio y encuadra el Analisis de Componentes Principales clásico dentro del caso particular de existencia de un solo grupo.

Dispersión Inducida, Componentes de Influencia

AMS: 62H25, 62H30.

1. Dispersión Inducida

Sea una muestra $m = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in I\!R^d$ cuyos elementos pertenecen a las clases $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_c\}$ y $t(\cdot)$ la función indicador que asocia a cada elemento $\mathbf{x} \in m$ la clase $\mathcal{C}_i$ a la que pertenece. Para cuantificar la influencia de cada grupo sobre $\mathbf{x}$ se define el concepto de Dispersión Inducida.

DEFINICIÓN: *Se denomina Función de Dispersión Inducida por la muestra m a una función contínua tal que a cada $\mathbf{x} \in I\!R^d$ le asocia una matriz $\Delta_{\mathbf{x}}$ simétrica y semidefinida positiva verificando que si $\mathbf{x} \to \boldsymbol{\mu}_i$ entonces $\Delta_{\mathbf{x}} \to \Sigma_i$, siendo $\boldsymbol{\mu}_i = E[\mathbf{X}/\mathcal{C}_i]$ y $\boldsymbol{\Sigma}_i = cov(\mathbf{X}/\mathcal{C}_i)$.*

En particular usando una combinación convexa de los Σ_i, puede construirse una función de dispersión inducida:

$$\Delta_{\mathbf{x}} = \sum_{i=1}^{c} \alpha_i^{\mathbf{x}} \boldsymbol{\Sigma}_i \quad \text{con} \quad \sum_{i=1}^{c} \alpha_i^{\mathbf{x}} = 1, \quad 0 \le \alpha_i^{\mathbf{x}} \le 1 \quad \text{y} \quad \boldsymbol{\Sigma}_i = \text{cov}(\mathbf{X}/C_i),$$

La matriz $\Delta_{\mathbf{x}}$, es una matriz virtual de dispersión, es simétrica y semidefinida positiva. A pesar de que $\Delta_{\mathbf{x}}$ no representa la dispersión de la muestra sí recoge la influencia que ejercen las clases sobre cada punto del espacio. Se plantearán distintas formulaciones de los $\alpha_i^{\mathbf{x}}$ consecuentes con las restricciones anteriores.

DEFINICIÓN: *Se denomina Transformación por Componentes de Influencia en $\mathbf{x}$ a la proyección $\mathbf{Y}_{\mathbf{x}} = \mathbf{X}\,\mathbf{U}_{\mathbf{x}}$, donde $\mathbf{U}_{\mathbf{x}} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_d]_{\mathbf{x}}$ es la matriz de autovectores de $\Delta_{\mathbf{x}}$.*

Como éstas varían de forma contínua de unos centroides a otros (por la continuidad de las transformaciones efectuadas) es de esperar que en cualquier punto $\mathbf{x}$ las Componentes de Influencia expliquen en parte la variabilidad de la muestra y en parte la influencia de los grupos. Se probará que las Componentes Principales son un caso particular de las CI.

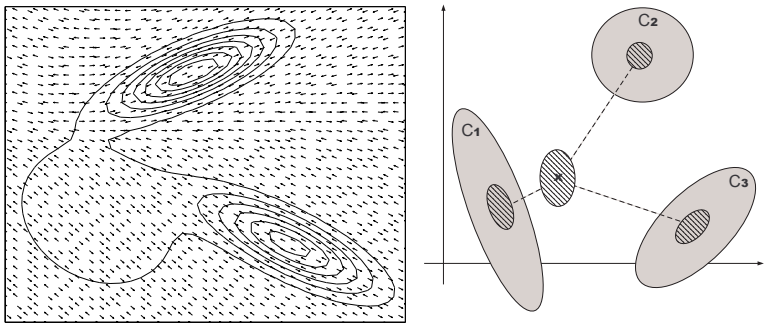

Figura 1: *Primer autovector de $\Delta_{\mathbf{x}}$ de una m.a.s. de una mixtura de tres distribuciones normales y formas adaptativas de las bolas en el método k-NN.*

1.1. Comportamiento de las CI

Se abordará el estudio de las CI en aquellos puntos del espacio donde $\Delta_{\mathbf{x}}$ manifieste alguna característica distintiva: a) zonas de máxima y mínima influencia de las clases, b) zonas de igual influencia de las clases, y c) zonas donde los autovectores de $\Delta_{\mathbf{x}}$ poseen capacidades explicativas análogas.

2. Variante del método k-NN empleando la metrica $\Delta_{\mathbf{x}}$

Se presentará una variante del método de clasificación de los k *vecinos mas cercanos* (k-NN) que denominaremos k-NN(Δ). La idea consiste en considerar la distancia asociada a la dispersión inducida, $d_{\mathbf{x}}(\mathbf{u},\mathbf{x}) = (\mathbf{u}-\mathbf{x})\Delta_{\mathbf{x}}^{-1}(\mathbf{u}-\mathbf{x})^T$, manteniendo inalterado el resto del algoritmo. Como puede observarse en la figura 1, con esta variante la métrica no es constante y la forma del entorno de un punto depende de la proximidad a cada clase.

3. Bibliografía

[1] Schott, J.R. (1997). Matrix Analysis for Statistics. Wiley & Sons.

[2] Marchette D. y Poston W. (1999). Local Dimensionality Reduction using Normal Mixtures. Computational Statistics, 14, 469-489, 1999.