

Modelos Jerárquicos Bayesianos de Clasificación: Aplicaciones a datos en alta dimensión

Luis Gutiérrez,*

Programa de Modelamiento y Medición en Salud, Escuela de Salud Pública, Facultad de Medicina, Universidad de Chile

Resumen

En la actualidad es común encontrar bases de datos en donde el número de muestras n es mucho menor que el número de variables o mediciones p , ($n \ll p$). Un ejemplo común donde ($n \ll p$) es en datos de espectroscopía. La espectroscopía mide valores de reflectancia o absorbancia de una sustancia o muestra bajo análisis en el espectro visible e infrarrojo cercano. Los valores de reflectancia o absorbancia son producidos por vibraciones en los enlaces químicos de la sustancia analizada. Los datos de espectroscopía se caracterizan por ser equiespaceados en función de la longitud de onda, muestran comportamientos no lineales y alta autocorrelación, lo que implica que la trayectoria de las curvas generalmente es suave.

Datos de espectroscopía han sido empleados con fines de clasificación utilizando propuestas de modelamiento basadas en proyecciones o selección de variables. Cuando se emplea selección de variables con fines de clasificación se corre el riesgo de perder información importante. La siguiente charla ilustra que es posible tratar los datos de espectroscopía como perfiles longitudinales. Bajo un enfoque de perfiles longitudinales la covarianza entre mediciones puede ser tratada con una estructura sencilla basada en unos pocos parámetros, simplificando el problema y pasando la dimensionalidad a un segundo plano. La propuesta se basa en un modelo lineal mixto completamente Bayesiano, donde la respuesta media se ajusta por splines penalizados incluyendo términos de efecto aleatorio para cada curva. La correlación serial se trata mediante un proceso estacionario autoregresivo. El modelo propuesto se aplica a un problema de autenticación de alimentos. Finalmente, se discutirá la importancia del desarrollo de modelos para datos de espectroscopía por el gran potencial que estos tienen en el diagnóstico de ciertos tipos de cancer.

Palabras clave: Splines penalizados; Autocorrelación serial; Autenticación de alimentos; Diagnóstico de cancer.

*E-mail:luisgutierrez@med.uchile.cl