

ANÁLISIS DISCRIMINANTE EN MODELOS RALOS: UNA APLICACIÓN A IMÁGENES
HIPERESPECTRALES

Expositor: Mery Lucia Picco (Universidad Nacional de Río Cuarto, mpicco@exa.unrc.edu.ar)
Autor/es: Mery Lucia Picco (Universidad Nacional de Río Cuarto, mpicco@exa.unrc.edu.ar);
Marcelo Ruiz (Universidad Nacional de Río Cuarto, mruiz@exa.unrc.edu.ar)

El análisis discriminante lineal (LDA en inglés) es una técnica simple y potente para predecir una respuesta cualitativa a partir de p variables predictoras con distribución normal multivariada. LDA consiste en particionar un vector p -dimensional en K clases a través de una proyección lineal cuyos parámetros son estimados a partir de un conjunto de n observaciones etiquetadas, llamado muestra de entrenamiento. En los problemas de alta dimensionalidad (p grande o incluso $p > n$) tanto LDA como muchos otros clasificadores enfrentan desafíos serios. Por ejemplo fallan en estimar modelos raros los cuales son importantes para alcanzar interpretabilidad y disminuir, posiblemente, el error de predicción. Clemmensen et al. (2011) proponen un clasificador denominado Sparse Discriminant Analysis (SDA) que formaliza la clasificación como un problema de regresión y estima los vectores discriminantes β_k que minimizan la suma de cuadrados de los residuos sujeta a dos constantes de ajuste, una de las cuales es una penalización de tipo *lasso* (norma ℓ_1). Así los vectores discriminantes estimados resultan raros. Por otro lado, Friedman et al. (2007) desarrollan un algoritmo para estimar la matriz de precisión rara usando una penalización ℓ_1 en la función de log-verosimilitud: $\log \det \Theta - \text{tr}(S\Theta) - \lambda \|\Theta\|_1$, donde Θ es la matriz de precisión, S la matriz de covarianza muestral y λ el parámetro de penalización. Usando esta estimación y el principio “plug-in” en el clasificador LDA, se obtiene un nuevo algoritmo, que llamaremos *LDA glasso*.

El objetivo de este trabajo es estudiar el desempeño de SDA y LDA glasso para la clasificación digital de imágenes de teledetección hiperespectrales, el cual es un problema de alta dimensionalidad. A fines comparativos se aplica también el algoritmo Random Forest (Breiman, 2001), técnica que no requiere supuesto distribucional para los datos y que ha demostrado un buen desempeño en los problemas clásicos de clasificación (Zou, 2018).

Palabras Claves: Aprendizaje estadístico, SDA, Glasso, imágenes hiperespectrales