Práctica 8: Modelos de regresión con un gran número de covariables (I)

Módulo de Modelos Lineales.
Máster de Bioestadística, Universitat de València.

Miguel A. Martinez-Beneito

Tareas

1. Repite la selección de variables que llevaras a cabo en la práctica anterior en la Tarea 1 apartado 2. En concreto, compara ahora los mejores modelos de las distintas dimensiones, empleando ahora los distintos tipos de validación cruzada que hemos introducido (LOOCV, 10-fold (por ejemplo) CV y conjunto de validación (2-fold CV)). Valora las diferencias que observas entre las técnicas de validación cruzada y las utilizadas en la práctica anterior, así como las diferencias obtenidas para las propias técnicas de validación entre sí.
2. El banco de datos `College` de la librería `ISLR` contiene datos (18 variables) de 777 universidades americanas. En concreto, estamos interesados en explicar el número de solicitudes de matriculación recibidas por estas universidades (`Apps`) como función del resto de variables del banco de datos.

- Divide la muestra en dos partes: un grupo train de 500 universidades y un grupo test con el resto. Elimina también de estos bancos de datos la primera variable (`Private`), por tratarse de una variable categórica
- Ajusta un modelo PCR, sobre el grupo train, eligiendo el número óptimo de covariables en el modelo mediante validación cruzada. Valora si resulta conveniente escalar las covariables del durante el proceso PCR
- Ajusta un modelo PLS, sobre el grupo train, eligiendo el número de covariables en el modelo mediante validación cruzada. Emplea, o no, el escalado de variables tal y como hicieras en el apartado anterior.
- Utiliza el grupo test para comparar ambos ajustes ¿por qué modelo te decidirías finalmente? Interpreta los resultados obtenidos ¿observas evidencia de sobreajuste para los modelos óptimos escogidos por ambos métodos?
- Por último, vamos a comparar, para el grupo test, el ajuste de los mejores modelos PCR y PLS con el mejor modelo según el procedimiento "best subset selection". Para ello, determina el modelo con mejor ajuste para el grupo train con el procedimiento "best subset selection". Utiliza C_p, por ejemplo, como criterio de comparación entre modelos de distintas dimensiones. Una vez calculado dicho modelo compara el MSE predictivo para el grupo test para este modelo y los mejores modelos PCR y PLS que hubieras determinado ¿Qué modelo te parece mejor opción atendiendo a estos resultados?