REGULARIZACIÓN Y REGRESIONES CON VARIABLES LATENTES

PRÁCTICA

Datos de Fearn (1983) y regresión *ridge*

Calibración de proteína a partir de un espectro Infrarrojo cercano (NIR). Este ejemplo procede de un artículo (Fearn 1983), en el que se afirmaba que eran inadecuados para regresión ridge. Posteriormente, en otro artículo (Hoerl *et al.* 1985) se corregían las afirmaciones del artículo anterior, y empleaban estos mismos datos para una regresión ridge. En el artículo original, los datos vienen subdivididos en dos partes, `Fearn.data.b.txt`, 24 observaciones, para conjunto de entrenamiento, y `Fearn.data.a.txt`, 26 observaciones para test.

En el script Fearn.01.r tenéis la entrada de datos y el ajuste de un modelo lineal a estos datos y, más abajo, una versión *a mano* de la regresión ridge. Podéis comprobar, con la exploración de los datos, que, efectivamente, estos datos son horribles para una regresión: las variables están muy correlacionadas y el número de condición de la matriz de la regresión es alto.

La función `lm.ridge` del paquete `MASS` tiene una implementación de la regresión ridge, ilustrada con los datos `longley`, un famoso dataset con notorio mal comportamiento en regresión. En el paquete `ElemStatLearn` también hay una función `simple.ridge`, ilustrada con el dataset `prostate`, que se encuentra en este mismo paquete. El paquete `glmnet` tiene una función general, de igual nombre, que permite ajustar tanto una regresión ridge como los otros métodos descritos más abajo.

El paquete `genridge` implementa un método gráfico de selección del parámetro de regularización en la regresión ridge.

Regresión en componentes principales y PLS

PCR (Principal Component Regression) y PLS (Partial Least Squares) son dos métodos de regresión sobre variables latentes, en los que se reemplazan las variables predictoras observadas por otras variables latentes, es decir, no observadas, construidas de forma que generen el mismo subespacio lineal, pero que sean ortonormales, por lo que tienen mejores propiedades numéricas. El paquete `pls` contiene funciones para PCR y para PLS.

El paquete `plsdof` contiene un estudio sobre los grados de libertad de la regresión PLS, según el artículo: Kraemer, Sugiyama (2011), *The Degrees of Freedom of Partial Least Squares Regression*. Aparte de implementaciones de PCR, PLS y regresión ridge, la función `benchmark.regression` permite comparar los tres métodos sobre unos datos.

Regresión *l1*

Las funciones `lqs`, `lmsreg`, `ltsreg` del paquete `MASS` (son tres nombres para una misma función). Implementan una variedad de regresión robusta. Otra variedad es la función `lmr`.

La función `rq` del paquete `quantreg` implementa la regresión *l1*.

Otra implementación de la regresión *l1* está en el paquete `pracma`. En el script `Advertising.r` teneis un ejemplo de su aplicación a los datos `Advertising`.

Finalmente, `L1pack` tiene la función `lad`, que también implementa la regresión *l1* ordinaria.

LASSO (Least Absolute Shrinkage and Selection Operator) y *elastic net*

El LASSO es muy parecido a la regresión ridge, pero la penalización es com la norma *l1* en vez de con la norma *l2*. En el paquete `lasso2` hay la implementación original del procedimiento LASSO. En el método *elastic net* tiene como penalización una suma de dos términos, uno con la norma l1 y

otro con la norma *l2*. En el paquete `elasticnet` hay la implementación del método de ese nombre (otro método de regresión regularizada). El paquete `glmnet` abarca las dos variantes, además de la regresión ridge, como se ha mencionado más arriba.