Tema 6
Técnicas de regresión avanzadas (II)
Técnicas multivariantes

Dr. Antoni Ferragut

unir
LA UNIVERSIDAD
EN INTERNET

- Veremos estrategias de regresión que permiten resolver problemas donde existan outliers o donde las variables predictoras presenten colinealidad:
- Regresión por mínimos cuadrados recortados (least trimmed squares LTS) para outliers (regresión robusta)
- Regresión penalizada (Ridge, LASSO, Red Elástica)

- Es un método robusto de estimación de los parámetros de una regresión lineal múltiple
- Escogemos un subcojunto de observaciones h de los n puntos originales cuyo ajuste por mínimos cuadrados posea la menor suma del cuadrado de los residuos
- Tomamos $n/2 \leq h \leq n$
- Inconveniente: el tiempo de computación crece demasiado al aumentar h

Aplicación:

- Escogemos de manera aleatoria el $h/n * 100$ % de los puntos y ajustamos por mínimos cuadrados.
- Repetimos *rep* veces el proceso y escogemos el subconjunto h^* que obtiene la mejor regresión.
- Reemplazamos un punto del subconjunto h^* por otro punto de n que no estuviese incluido en h^*. Repetimos el proceso *rep2* veces.

Situaciones:

- Alternativa consistente para realizar selección de variables
- Útil cuando hay colinealidad o cuando hay muchas variables predictoras
- Se añade una restricción sobre el tamaño de los coeficientes de regresión

- Se busca minimizar el error cuadrático medio con la restricción $\sum_{i=1}^{p} \beta_i^2 < s$, donde s es el parámetro de ajuste.
- No se produce selección de variables, sino que los coeficientes se encogen a medida que s disminuye (más restricción).
- Es recomendable normalizar (estandarizar) las variables, ya que la escala influye en la estimación de los coeficientes.

Ventajas:

- Se reduce la varianza para distintas muestras (a costa de aumentar el sesgo). De esta manera los coeficientes β no varían drásticamente frente a cambios en la muestra
- Cuando existen dos variables predictoras muy correlacionadas, su peso en la regresión se divide, y por tanto sus coeficientes son similares.

- Es similar a la regresión Ridge, pero con la restricción $\sum_{i=1}^{p} |\beta_i| < s$.
- Los coeficientes β pueden encogerse hasta 0, por tanto es posible hacer selección de variables.
- Cuando algunas variables predictoras están muy correlacionadas puede seleccionar solamente una y descartar el resto (la varianza aumenta).

- Intervienen al mismo tiempo las penalizaciones de tipo LASSO y de tipo Ridge.
- Tenemos, a partir de un hiperparámetro $r \in [0, 1]$, una combinación de las dos:

$$(1 - r) \sum_{i=1}^{p} \beta_i^2 + r \sum_{i=1}^{p} |\beta_i| < s.$$

- r puede fijarse de antemano, aunque en ocasiones es más adecuado ajustarlo mediante técnicas de validación.