UNIVERSITÉ DU LUXEMBOURG

High Dimensional Statistics
2024

Mark Podolskij
mark.podolskij@uni.lu

HOMEWORK 3

Exercise 1. Consider the Ridge-regression estimator, obtained replacing the l_1 penalty in the definition of the Lasso estimator with an l_2 penalty :

$$\widehat{b}_{\text{Ridge}}(\lambda) := \arg\min_{b\in\mathbb{R}^k} \left(\frac{\|Y - Xb\|_2^2}{n} + \lambda \|b\|_2^2 \right).$$

Determine the explicit formula for $\widehat{b}_{\text{Ridge}}(\lambda)$.

Exercise 2. Let X and Y be two random variables. Show that $\text{corr}(X, Y) = 1$ if and only if there exist $a > 0, b \in \mathbb{R}$ such that $X = aY + b$ $\mathbb{P}$-almost surely.
Hint : If a random variable Z is such that $\text{var}(Z) = 0$, then $Z = c$ $\mathbb{P}$-almost surely, for some constant $c \in \mathbb{R}$.

Exercise 3. Assume that the data X is distributed according to probability measure $\mathbb{P}_\theta$, where $\theta \in \Theta$ (Θ is the parameter space). Let θ_0 be the true parameter and $\widehat{\theta}_{\text{ML}}$ the maximum likelihood estimator of θ_0. Show : If $f : \Theta \to \tilde{\Theta}$ a bijective function then $f(\widehat{\theta}_{\text{ML}})$ is the maximum likelihood estimator of $f(\theta_0)$.

Exercise 4. (Numerical exercise to solve with R or Python).
The goal of an experimental research is to analyse the link between the value of the specific prostate antigen and some covariates in subjects undergoing prostatectomy surgery.
The prostate dataset can be found at the link : "https ://hastie.su.domains/ElemStatLearn/" under "Data".

1. Build the regression model for the variable prostate antigen (lpsa) : $Y_i = b_0 + \sum_{j=1}^{8} b_j t_{ij} + \epsilon_i$ and estimate b_0 and b_j, for $j \in \{1, ..., 8\}$.
2. Compute the Lasso estimator and plot them the estimated coefficients.
   Hint : a useful function can be found in the R package "lasso2".
3. Report two values for λ : "lambda.min" and "lambda.1se", where “lambda.min” is the λ at which the smallest mean squared error (MSE) is achieved and “lambda.1se” is the largest λ at which the MSE is within one standard error of the smallest MSE (default). These quantities can be computed using the R package "glmnet".
   Report the number of nonzero coefficients for the selected values of λ and the corresponding estimated coefficients.