Introduction Supervised Learning
IMA205

Pietro Gori

Deadline: Upload the answers and the two notebooks as a single .zip file to the site pédagogique before the 5th of February 2019 (23h59). Name it as 'TP1-IMA205-YOUR-SURNAME.zip'.

Theoretical questions

OLS

We have seen that the OLS estimator is equal to $\beta^* = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$ which can be rewritten as $\beta^* = H\mathbf{y}$. Let $\tilde{\beta} = C\mathbf{y}$ be another linear unbiased estimator of β where C is a $d \times n$ matrix, e.g. $C = H + D$ where D is a non-zero matrix .

- Demonstrate that OLS is the estimator will the smallest variance: compute $\mathbf{E}[\tilde{\beta}]$ and $\text{Var}(\tilde{\beta}) = \mathbf{E}[(\tilde{\beta} - \mathbf{E}[\tilde{\beta}])(\tilde{\beta} - \mathbf{E}[\tilde{\beta}])^T]$ and show when and why $\text{Var}(\beta^*) < \text{Var}(\tilde{\beta})$. Which assumption of OLS do we need to use ?

Ridge regression

Suppose that both $\mathbf{y}$ and the columns of $\mathbf{x}$ are centered ($\mathbf{y}_c$ and $\mathbf{x}_c$) so that we do not need the intercept β_0. In this case, the matrix $\mathbf{x}_c$ has d (rather than d+1) columns. We can thus write the criterion for ridge regression as:

$$\beta^*_{ridge} = \arg\min_{\beta}(\mathbf{y}_c - \mathbf{x}_c\beta)^T(\mathbf{y}_c - \mathbf{x}_c\beta) + \lambda||\beta||_2^2 \tag{1}$$

- Show that the estimator of ridge regression is biased (that is $\mathbf{E}[\beta^*_{ridge}] \neq \beta$).
- Recall that the SVD decomposition is $\mathbf{x}_c = UDV^T$. Write down by hand the solution β^*_{ridge} using the SVD decomposition. When is it useful using this decomposition ? Hint: do you need to invert a matrix ?
- Remember that $\text{Var}(\beta^*_{OLS}) = \sigma^2(\mathbf{x}^T\mathbf{x})^{-1}$. Show that $\text{Var}(\beta^*_{OLS}) \geq \text{Var}(\beta^*_{ridge})$.
- When λ increases what happens to the bias and to the variance ? Hint: Compute MSE = $\mathbf{E}[(y_0 - x_0^T\beta^*_{ridge})^2]$ at the test point (x_0, y_0) with $y_0 = x_0^T\beta + \epsilon_0$ being the true model and $x_0^T\beta^*_{ridge}$ the ridge estimate.
- Show that $\beta^*_{ridge} = \frac{\beta^*_{OLS}}{1+\lambda}$ when $\mathbf{x}_c^T\mathbf{x}_c = I_d$

Elastic Net

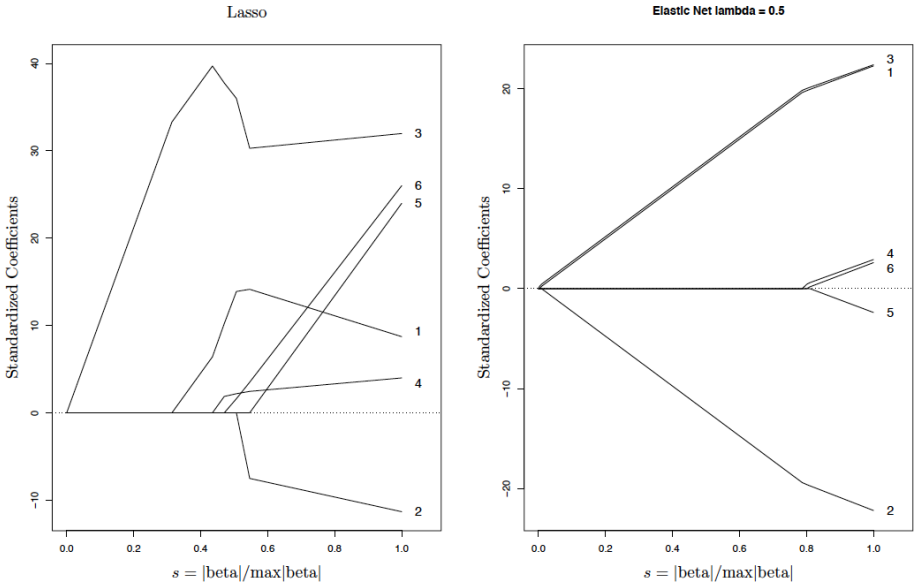
Using the previous notation, we can also combine Ridge and Lasso in the so-called Elastic Net regularization :

$$\beta^*_{ElNet} = \arg\min_{\beta}(\mathbf{y}_c - \mathbf{x}_c\beta)^T(\mathbf{y}_c - \mathbf{x}_c\beta) + \lambda_2||\beta||_2^2 + \lambda_1||\beta||_1 \tag{2}$$

Calling $\alpha = \frac{\lambda_2}{\lambda_1+\lambda_2}$, solving the previous Eq. is equivalent to:

$$\beta^*_{ElNet} = \arg\min_{\beta}(\mathbf{y}_c - \mathbf{x}_c\beta)^T(\mathbf{y}_c - \mathbf{x}_c\beta) + \lambda\left[\alpha(\sum_{j=1}^{d}\beta_j^2) + (1-\alpha)(\sum_{j=1}^{d}|\beta_j|)\right] \tag{3}$$

- This regularization overcomes some of the limitations of the Lasso, notably:
  - If $d > N$ Lasso can select at most N variables → ElNet removes this limitation
  - If a group of variables are highly correlated, Lasso randomly selects only one variable → with ElNet correlated variables have a similar value (grouped)
  - Lasso solution paths tend to vary quite drastically → ElNet regularizes the paths
  - If $N > d$ and there is high correlation between the variables, Ridge tends to have a better performance in prediction → ElNet combines Ridge and Lasso to have better (or similar) prediction accuracy with less (or more grouped) variables



- Compute by hand the solution of Eq.2 supposing that $\mathbf{x}_c^T\mathbf{x}_c = I_d$ and show that the solution is: $\beta^*_{ElNet} = \frac{(\beta^*_{OLS})_j \pm \frac{\lambda_1}{2}}{1+\lambda_2}$

LDA

In LDA we assume that all classes have the same covariance matrix Σ.

- What happens if we assume that each class k has its own covariance matrix Σ_k? Compute $f^*(x_j)$ given a training set $\mathcal{T}$ where x_j is a test sample.
- Show that the solution is a quadratic discriminant function. This is called Quadratic Discriminant Analysis (QDA).