Université Grenoble Alpes
Model selection for large scale learning
emilie.devijver@univ-grenoble-alpes.fr

Lab 2: variable selection

We would like to have an illustration of the results get during the lectures. You can use your favorite language (R or Python are recommended).

For that, we propose the following exercises (one is mandatory, choose the one that you like the much, and *of course* you can do more ;))

Send the report before Friday December 11, 5pm. You can provide a report by team if you have worked by team (less than 3 people ...).

1 Illustration of Lasso and Ridge, and selection by CV and AIC

1. Construct a function to generate n observations drawn from a linear Gaussian model, with $p = 200$ covariates, distributed with respect to a Gaussian with mean $\mathbf{0}_{200}$ and covariance I_{200}, β sparse with nonzero coefficients indexed by 1 to 20, drawn randomly from a uniform distribution on $[1, 2]$ and Y generated from this linear regression model with a Gaussian noise with variance 0.1.

2. Consider the following 4 estimators: (to implement or to use directly)

   - The LSE
   - Use the Ridge estimator (to use directly), defined by

$$\hat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda) = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

   - Use the Lasso estimator, defined by

$$\hat{\boldsymbol{\beta}}^{\text{lasso}}(\lambda) = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

3. Compute each estimator (or sequence of estimators for the Ridge or the Lasso) for $n \in \{100, 200, 500\}$. Compare with the LSE for $n \in \{201, 500\}$.

4. For each estimator (or sequence of estimator), for each size of sample, compute the empirical risk (on a new sample of size 1000) for each collection of models, with respect to the least square loss.

5. Select a model in the Lasso collection and in the Ridge collection using the CV and the AIC.

6. Summarize the results via boxplots and conclude.

2 Illustration of theoretical results for the Lasso

We want to analyze numerically the theoretical results get for the Lasso (in estimation, prediction and support estimation).

1. First fix $p = 500$ and $n = 100$. Generate 3 datasets for which the constants γ defined in the mutual incoherence are different: close to 1, close to 0, and close to 0.5. Compute numerically those values. Generate also a dataset where a coefficient of β is really small, such that the last theorem about support recovery will lead to the wrong support.

2. For those datasets, test the performance of the Lasso for 1/ variable selection, 2/ prediction and 3/ estimation.

3. Have a look on the regularization parameter selected by cross-validation for the severals datasets. What is its order? For each dataset, test several dimensions: $p = 100$ and $n = 100$, $p = 500$ and $n = 100$, $p = 10000$ and $n = 100$. Compare with the performances of the theoretic regularization parameter: $\lambda = 2\sigma\sqrt{3\log(p)/n}$.

3 Open question

We would like to test what to do on real data.

Two datasets can be considered: `Prostate`, which have 97 observations of 9 covariates, and `NCI60`, with 64 observations of 6830 covariates.

Propose the best model, up to you, among those we have seen together.

Kind reminder:

- Preprocessing: centering, rescaling, screening
- Different estimators: LSE, Lasso, Ridge, Elastic-Net
- Several criteria: train error, test error, interpretability
- How to tune parameters: Hold-Out, Cross Validation (K=1, 5, 10, n), AIC, BIC