ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

D-BSSE
Department of Biosystems
Science and Engineering

Statistical Models in Computational Biology

Niko Beerenwinkel
Auguste Rimaite
Rudolf Schill
David Dreifuss

Due 16th of May 2024
Please submit your project with the filename `Lastname(s)_Project10.pdf`.

Problem 27: Uniqueness of predictions from the lasso **(2 points)**

Given any response vector $\mathbf{y}$, input matrix $\mathbf{X}$ and regularization parameter $\lambda \geq 0$, suppose we have two lasso solutions $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ such that

$$\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\hat{\beta}^{(1)}\right\|_2^2+\lambda\left\|\hat{\beta}^{(1)}\right\|_1=\frac{1}{2}\left\|\mathbf{y}-\mathbf{X}\hat{\beta}^{(2)}\right\|_2^2+\lambda\left\|\hat{\beta}^{(2)}\right\|_1=c^*$$

In general, the lasso criterion is convex and since the solution set of a convex minimization problem is convex, we have $\alpha\hat{\beta}^{(1)}+(1-\alpha)\hat{\beta}^{(2)}$ also in the solution set for any $\alpha \in (0,1)$, resulting in uncountably many lasso solutions.

Show that $\mathbf{X}\hat{\beta}^{(1)}=\mathbf{X}\hat{\beta}^{(2)}$, i.e. $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$ give the same predictions.
(*hint: Given a convex set S, a function $f: S \to \mathbb{R}$ is said to be strictly convex if*

$$\forall s_1 \neq s_2 \in S, \forall \alpha \in (0,1): \quad f(\alpha s_1+(1-\alpha)s_2) < \alpha f(s_1)+(1-\alpha)f(s_2)$$

Use the strict convexity of the loss function $f(u)=\|y-u\|_2^2$ and convexity of the l_1 norm to establish a contradiction.)

Problem 28: Regularization and Bayesian regression model **(3 points)**

Consider the linear regression model

$$y=X\beta+\varepsilon, \quad \varepsilon \sim N(0,\sigma^2 I_n),$$

where y is an $n \times 1$ vector of the dependent variables, X is the $n \times p$ design matrix, β is the $p \times 1$ vector of the regression coefficients, and ε is the $n \times 1$ vector of errors.
a) In frequentist statistics, the Ridge regression coefficients are chosen as minimizers of

$$\|y-X\beta\|^2+\lambda\|\beta\|^2, \quad \lambda \geq 0.$$

Show that for some λ, the Ridge regression coefficients are equivalent to the maximum a posteriori (MAP) estimator, if we assume a normal prior for the coefficients

$$\beta \sim N(0,\sigma_\beta^2 I_p).$$

b) The (frequentist) Lasso estimates are defined as minimizers of

$$\|y-X\beta\|^2+\lambda\|\beta\|_1,$$

where

$$\|\beta\|_1=\sum_{k=1}^{p}|\beta_k|.$$

Show that for some λ, the Lasso coefficients are equivalent to the MAP estimator if we assume Laplace priors for the coefficients

$$\pi(\beta) = \prod_{k=1}^{p} \frac{1}{2b} e^{-\frac{|\beta_k|}{b}}.$$

Problem 29: Variable selection under various norms **(5 points)**

Solve this exercise in R. Use the `caret` package for data construction and `glmnet` and `pROC` packages for model fitting and performance evaluation.

The `yeastStorey.rda` data frame contains marker and gene expression information of 112 F1 segregants derived from a yeast genetic cross of two strains. The first column is a binary marker (response) denoting presence (1) or absence (0) of a SNP and the remaining columns correspond to the gene expression values across the segregants (predictors).

In this task, you will fit a logistic regression model which assumes a linear relationship between the predictors and the dependent variable in the log-odds scale. In logistic regression, the likelihood function, which is maximised during the model fitting, is derived from the binomial distribution. Binary logistic regression predicts the probability of the observation falling into one of two categories (such as the presence or absence of SNP in this case), based on the values of the predictors.

1. Load the data and construct the design matrix $\mathbf{X}$ and response variable $\mathbf{y}$, respectively. Randomly split the data into training set (70%) and test set (30%). For reproducibility set a seed in the beginning. (1 point)

2. Using 10-fold cross-validation, find the optimum λ and optimum α using elastic-net model on the training set. For binary response variables you need to call `cv.glmnet` with `family = "binomial"`. In order to reduce computation time, restrict the search space of α to $\{0, 0.1, 0.2, \cdots, 1\}$. For the optimal α, plot the mean cross-validated error as a function of log λ and the trace curve of coefficients as a function of log λ. (2 points)

3. Fit the final model with optimal α and optimal λ on the training set using `glmnet` and predict the response on the test dataset. Report the variables selected, plot the ROC curve and report the corresponding AUC (area under the curve). (2 points)