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
Pedro Ferreira
Xiang Ge Luo
David Dreifuss

Due 19th of May 2022
Please submit your project with the filename `Lastname(s)_Project10.pdf`.

Problem 27: Uniqueness of predictions from the lasso **(3 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.

1. 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. (2 points)
   (*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.)
2. If $\lambda > 0$, show that $\left\|\hat{\beta}^{(1)}\right\|_1=\left\|\hat{\beta}^{(2)}\right\|_1$ (1 point)

Problem 28: Bayesian priors as regularizers **(2 points)**
A linear regression problem can also be approached with a Bayesian perspective, by adding a prior for the parameter vector β. Consider the Lasso estimator

$$\hat{\beta}^{lasso}=\text{argmin}_\beta\frac{1}{2}\|\mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1$$

show that, for some λ and some b, $\hat{\beta}^{lasso}$ is equivalent to the Maximum a Posteriori (MAP) estimate $\hat{\beta}^{MAP}$ of the Bayesian linear regression with a Laplace prior on β. The Laplace prior has the form:

$$\pi(\beta)=\prod_{j=1}^{p}\frac{1}{2b}\exp\{-|\beta_j|/b\}$$

(hint: The MAP estimate in Bayesian linear regression is obtained by optimizing the posterior or log-posterior, instead of the likelihood or log-likelihood.)

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).

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 the seed to 42 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. Fit the final model with the optimal parameters on the training set. For reducing computation time restrict the search space of α to $\{0, 0.1, 0.2, \cdots, 1\}$. (2 points)
3. Predict the response on the test dataset using the final model. Plot the cross-validation error as a function of $\log \lambda$, trace curve of coefficients as a function of $\log \lambda$, and the ROC curve for the optimal α. Lastly, report the corresponding AUC (area under the curve) of the ROC curve and the variables selected. (2 points)