Topics in Statistical Sciences 2 – Exam exercise 2

Søren Højsgaard and Torben Tvedebrink

Version: 03/11/2017 11:15

This exercise is about the *Penalised Regression* as discussed in lectures 4 – 6 of Topics in Statistical Sciences 2. During the oral exam you will have 20 min to present the exercise. You decide what topics to cover and how to present them, however, we will ask questions to any part of the covered curricula, exercise and presentation.

1. Show that augmenting the response, $\boldsymbol{y}$, and design matrix, $\mathbf{X}$, appropriately, one can obtain the ridge regression estimator from the OLS expression, $(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top\tilde{\boldsymbol{y}}$, where $\tilde{\boldsymbol{y}}$ and $\tilde{\mathbf{X}}$ are the augmented response and design matrix, respectively.

In Section 5.4.3 "Logistic Regression and Generalized Linear Models" of SLS the authors discuss how to fit the LASSO penalised logistic regression model using coordinate descent.

This relies on the approximation of the log-likelihood using a second-order Taylor expansion, which in turn leads to the re-weigthed least squares method for GLM likelihoods.

Hence, we have that Equation (5.54) in SLS expresses the log-likelihood

$$\begin{aligned}\ell(\beta_0,\beta) &= \frac{1}{n}\sum_{i=1}^{n}\Big[y_i(\beta_0+x_i^\top\beta)-\log\{1+\exp(\beta_0+x_i^\top\beta)\}\Big] \qquad \text{(SLS: 5.54)}\\ &= -\frac{1}{2n}\sum_{i=1}^{n} w_i^*(z_i^*-\beta_0-x_i^\top\beta)^2 + C_2(\beta_0-\beta_0^*,\beta-\beta^*)\\ &= \ell_Q(\beta_0,\beta)+C_2(\beta_0-\beta_0^*,\beta-\beta^*),\end{aligned}$$

where the C_2-term is the error committed when discarding higher order terms in the Taylor expansion, and (β_0^*,β^*) are the current parameter estimates. The z_i^* and w_i^* are referred to as the working response, $z_i^*=\beta_0^*+x_i^\top\beta^*+[y_i-p^*(x_i)]/w_i^*$, and weights, $w_i^*=p^*(x_i)[1-p^*(x_i)]$, respectively, with $p^*(x_i)$ being the current estimate of the probability that $y_i=1$ given x_i, $p^*(x_i)=\exp(\beta_0^*+x_i^\top\beta^*)/[1+\exp(\beta_0^*+x_i^\top\beta^*)]$.

2. Show how to obtain $\ell_Q(\beta_0,\beta)$ from (SLS: 5.54).

Hint: Data Mining Self-study Exercise 9 on Iterative Least Squares for logistic regression

The elastic net for logistic regression is given by solving Equation (SLS: 5.54) with the $\lambda P_\alpha(\beta)$ penalty added, $P_\alpha(\beta)=\sum_{j=1}^{p}\{(1-\alpha)\beta_j^2/2+\alpha|\beta_j|\}$. However, this expression is not directly applicable in coordinate descent. Therefore, the SLS-authors use coordinate descent on the ℓ_Q expression.

3. Derive the update equations for β_j, based on $-\ell_Q+\lambda P_\alpha(\beta)$ for $j=1,\ldots,p$.

In the `leukemia.csv` (on moodle), gene-expression measurements on $p=3571$ genes measured on blood samples from $n=72$ leukemia patients. They were classified into two classes, 47 acute lymphoblastic leukemia (`ALL`) and 25 myeloid leukemia (`AML`).

4. Analyse the data to identify the relevant predictors for the two classes. Your analysis should include selection of λ, estimates of prediction accuracy, parameter variability, etc.