**Statistical and Machine Learning** **Homework assignment 2**
**Grün** **WS 2023/24**

# 2 Regularized generalized linear models

**Exercise 1:**

Consider $\hat{\beta}^{\text{ridge}}$ and $\hat{\beta}^c$ determined by solving the two optimization problems

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

$$\hat{\beta}^c = \arg\min_{\beta^c} \left\{ \sum_{i=1}^{N} (y_i - \bar{y} - \beta_0^c - \sum_{j=1}^{p} (x_{ij} - \bar{x}_j)\beta_j^c)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\},$$

where $\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$.

Specify how $\hat{\beta}_j^{\text{ridge}}$ is related to $\hat{\beta}_j^c$, $j = 0, 1, \ldots, p$.

**Exercise 2:**

Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau^2 \boldsymbol{I})$, and Gaussian sampling model $y \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I})$. Find the relationship between the regularization parameter $\lambda$ in the ridge formula, and the variances $\tau^2$ and $\sigma^2$.

**Exercise 3:**

Show that in case the design matrix $\boldsymbol{X}$ of a linear regression model is orthonormal (i.e., $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$), that the estimators of $\beta_j$ are given by the following equations with $\hat{\beta}_j$ denoting the ordinary least squares estimate:

(a) Best subset of size $M$:

$$\hat{\beta}_j I(|\hat{\beta}_j| \geq |\hat{\beta}_M|).$$

(b) Ridge with penalty $\lambda$:

$$\hat{\beta}_j/(1 + \lambda).$$

(c) Lasso with penalty $\lambda$:

$$\text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+.$$

$I()$ denotes the indicator function, sign() the sign of its argument ($\pm 1$) and $x_+$ the "positive part" of $x$.

**Exercise 4:**

Suppose $y_i$ has a Poisson distribution with $g(\mu_i) = \beta_0 + \beta_1 x_i$, where $x_i = 1$ for $i = 1, \ldots, n_A$ from group A and $x_i = 0$ for $i = n_A + 1, \ldots, n_A + n_B$ from group $B$, and with all observations being independent. Show that for the log-link function, the GLM likelihood equations imply that the fitted means $\hat{\mu}_A$ and $\hat{\mu}_B$ equal the sample means.

**Exercise 5:**

(a) Show how the binomial distribution with success probability $\pi$ and number of trials value $T$ can be written as a univariate exponential dispersion family given by

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right).$$

(b) Write down the log-likelihood for a binomial regression model with logit link for a sample $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, N$, where $y_i$ denotes the number of successes out of $T_i$ trials.

(c) Derive the score function $\boldsymbol{s}(\beta)$, i.e., the derivative of the log-likelihood function with respect to $\beta$.

**Exercise 6:**

Use artificial data to perform LASSO and ridge regression. Set a random seed before the analysis.

(a) Draw 100 observations from a 100-dimensional standard multivariate normal distribution. This is the matrix of covariates `X` of dimension $100 \times 100$.

(b) Draw 100 observations for the dependent variable given by

$$y = \sum_{i=1}^{10} x_i + \epsilon,$$

with $\epsilon \sim N(0, 0.1)$.

(c) Fit LASSO and ridge models with different values of $\lambda$ using function `glmnet` from package **glmnet**.

*Note:* Note that the default is `intercept = TRUE`. Keep this default to fit a model including an intercept to `X` and `y`.

(d) Create the default plots for the returned objects and interpret them. Create also the plots where the argument `xvar` is set to `"lambda"` and interpret them. Point out the specific differences between the solutions obtained for LASSO and ridge regression.

(e) Determine the number of non-zero coefficients and the model fit as measured by the `deviance()` (= RSS) in dependence of $\lambda$ for LASSO and ridge. Visualize these results and comment on them.

*Note:* You can use `predict(fit, type = "nonzero")` to obtain the indices of variables with non-zero coefficients for the `fit` object returned by `glmnet()` which contains the $\lambda$ sequence used in `fit$lambda`.

**Exercise 7:**

Assume the following data generating process where for each observation first $G \in \{1, \ldots, K\}$ is drawn from a multinomial distribution with parameter $\boldsymbol{\pi}$ indicating class membership:

$$G \sim \text{Multinomial}(\boldsymbol{\pi})$$

and conditional on $G = k$ the vector $\boldsymbol{x} \in \mathbb{R}^p$ is drawn from a multivariate normal distribution with mean parameter $\boldsymbol{\mu}_k$ and variance-covariance matrix $\boldsymbol{\Sigma}_k$:

$$\boldsymbol{x}|G = k \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

- Determine

$$\log\left(\frac{\Pr(G = k|\boldsymbol{x})}{\Pr(G = l|\boldsymbol{x})}\right)$$

and show that it is a quadratic function in $\boldsymbol{x}$.

- Assume $\mathbf{\Sigma}_k \equiv \mathbf{\Sigma}$ for all $k = 1, \ldots, K$ and show that in this case

$$\log\left(\frac{\Pr(G = k|\boldsymbol{x})}{\Pr(G = l|\boldsymbol{x})}\right)$$

  is a linear function in $\boldsymbol{x}$.

  Comment on how these log odds change in the univariate case if $\mu_k$ is changed in comparison to $\mu_l$, if the variance $\sigma^2$ is increased and the group size $\pi_k$ is increased in comparison to $\pi_l$.

**Exercise 8:**

Assume the following data generating process:

$$\begin{aligned} G &\sim \text{Multinomial}((0.5, 0.5)) \\ x|G = 1 &\sim N(0, 1) \\ x|G = 2 &\sim N(\mu, \sigma^2). \end{aligned}$$

(a) Determine the optimal decision boundary, i.e., where $\Pr(G = 1|x) = \Pr(G = 2|x)$.

(b) Calculate the expected misclassification error if the optimal decision boundary is used to classify observations and if:

  (a) $\mu = 0$, $\sigma^2 = 2$.

  (b) $\mu = 1$, $\sigma^2 = 1$.

  This error is also referred to as *Bayes rate* and characterizes the difficulty of the problem.

(c) Visualize the class-specific densities and the boundaries.

**Exercise 9:**

The dataset `icu` in package **aplore3** contains information on patients who were admitted to an adult intensive care unit (ICU). The aim is to develop a predictive model for the probability of survival to hospital discharge of these patients.

(a) Fit a logistic regression model using all potentially useful covariates as regressors. Inspect the estimated coefficients.

(b) Assess if complete or quasi-complete separation is a problem and transform categorical variables to have less categories to alleviate this problem. Note that coefficients which are large in absolute terms might indicate complete separation. Refit the logistic regression model to the resulting data set.

(c) Binarize the variable `loc` by combining the levels `"Deep stupor"` and `"Coma"` and the variable `race` by combining the levels `"Black"` and `"Other"`.

(d) Use stepwise procedures to estimate a suitable model based on AIC and BIC. Function `step()` has an argument `k` to specify the penalty, which implies by default to use AIC. Compare the selected models.

(e) Compare the in-sample log-likelihoods and the in-sample misclassification rates of the full model to those selected with AIC and BIC.

**Exercise 10:**

Use the `Default` data set included in package **ISLR2** to fit a logistic regression to predict the probability of default using income and balance. Set a random seed before beginning your analysis.

(a) Fit a logistic regression model that uses income and balance to predict default.

(b) Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:

   i. Split the sample set into a training set and a validation set.

   ii. Fit a multiple logistic regression model using only the training observations.

   iii. Obtain a prediction of default status for each individual in the validation set by computing the predicted probability of default for that individual, and classifying the individual to the default category if the predicted probability is greater than 0.5.

   iv. Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.

(c) Repeat the process in (b) 100 times, using 100 different splits of the observations into a training set and a validation set. Assess how the misclassification rate and the false negative rate on the validation set vary with the proportion of defaults in the validation set and comment on the results obtained.

*Hint:* The predicted probabilities can be obtained for a fitted logistic regression model `glmfit` using `predict(glmfit, type = "response")` where the `newdata` argument can be used to provide the test data.