Statistical and Machine Learning **Exercise 2**
Grün, Bajons **WS 2025/26**

2 Regularized generalized linear models

Sub-task 1:

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

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

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

  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 $\hat{\beta}_j^c$, $j = 0, 1, \ldots, p$.

- Show that the ridge regression estimator 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 $\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I})$. Find the relationship between the regularization parameter λ in the ridge formula, and the variances τ^2 and σ^2.

Sub-task 2:

Show that in case the design matrix $\boldsymbol{X}$ of a linear regression fulfills $\frac{1}{N}\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{I}$, that the estimators of β_j are given by the following equations with $\hat{\beta}_j$ denoting the ordinary least squares estimator:

(a) Best subset of size M:

$$\hat{\beta}_j I(\text{rank}(|\hat{\beta}_j|) \leq M).$$

(b) Ridge with penalty λ:

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

(c) Lasso with penalty λ:

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

rank() gives the ranks for the absolute values of the OLS estimates with rank 1 for the largest value, $I()$ denotes the indicator function, sign() the sign of its argument (± 1) and x_+ the "positive part" of x.

Hint: Note that in the case of an orthonormal $\frac{1}{\sqrt{N}}\boldsymbol{X}$, the OLS estimator of β is given by

$$\hat{\beta} = \frac{1}{N}\boldsymbol{X}^\top \boldsymbol{y},$$

and the columns of $\boldsymbol{X}$ form a p-dimensional orthogonal basis for the column space of $\boldsymbol{X}$.

Sub-task 3:

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 generalized linear model (GLM) likelihood equations imply that the fitted means $\hat{\mu}_A$ and $\hat{\mu}_B$ equal the sample means.

Sub-task 4:

(a) Show how the binomial distribution with success probability π 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 β.

Sub-task 5:

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 100×100.

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

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

with $\epsilon \sim N(0, 0.1)$. This implies that y only depends on the first ten x variables and the variance of the noise is given by 0.1, i.e., the standard deviation is $\sqrt{0.1}$.

(c) Fit lasso and ridge models with different values of λ using function `glmnet` from package **glmnet**.

Note: The default in `glmnet` 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 λ 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 λ sequence used in `fit$lambda`.

Sub-task 6:

Assume the following data generating process where first a group indicator $G \in \{1, 2\}$ is drawn and then x conditional on the value of G using:

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

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

$$g(x) = \log\left(\frac{\Pr(G = 1|x)}{\Pr(G = 2|x)}\right)$$

and determining the roots of $g(x)$. Determine then the regions for x where $g(x)$ is positive and where it is negative.

(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.

Hint: The expected misclassification error is given by

$$\pi_1 \int_{-\infty}^{\infty} I(g(x) < 0) f(x|G = 1) dx + \pi_2 \int_{-\infty}^{\infty} I(g(x) > 0) f(x|G = 2) dx,$$

where $I()$ is the indicator function and (π_1, π_2) correspond to the success probabilities of the groups, i.e., they are equal to $(0.5, 0.5)$ in this example.

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

Sub-task 7:

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.

Sub-task 8:

Use the `Bikeshare` data set included in package **ISLR2** to fit a regression model to predict the number of bikers (`bikers`) in dependence of `mnth`, `hr`, `workingday`, `weathersit` and `temp`.

(a) Summarize the dataset considering only the variables needed for the regression.

Transform `workingday` to create a categorical variable and `weathersit` to combine the two levels containing `"rain/snow"` (`"light rain/snow"`, `"heavy rain/snow"`) in one level together.

(b) Fit the following three regression models using as covariates main effects for `month`, `hr`, `workingday`, `weathersit` and `temp`:

- a linear regression model to predict the total number of bikers (`bikers`);
- a linear regression model to predict the logarithmized total number of bikers (log(`bikers`));
- a Poisson regression model with log-link to predict the total number of bikers (`bikers`).

(c) Extract the regression coefficients for `temp` for the three models, interpret the estimated coefficients and compare the results.

(d) Compare the model fit by assessing

- the Poisson log-likelihood based on the predicted number of bikers;
- determining the mean squared error between the predicted and observed number of bikers.

Assess if there are any issues evaluating these criteria and compare the performance.