

## 1 Introduction and penalized regression

### Sub-task 1:

Consider a balanced regression problem where for each input  $\mathbf{x}_i$ ,  $i = 1, \dots, N$ , one has  $J$  repeated outputs  $y_{ij}$ ,  $j = 1, \dots, J$  and one fits a parameterized model  $f_\theta(\mathbf{x})$  by least squares.

- Show that the fit can be obtained from a least squares problem involving only  $\mathbf{x}_i$  and the average values  $\bar{y}_i = \frac{1}{J} \sum_{j=1}^J y_{ij}$ .
- Explain how the least squares problem changes if the design is not balanced, i.e., one has different number of repetitions for each input  $\mathbf{x}_i$ .

### Sub-task 2:

Assume a set of training data  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$  drawn at random from a population as well as some test data  $(\tilde{\mathbf{x}}_1, \tilde{y}_1), \dots, (\tilde{\mathbf{x}}_M, \tilde{y}_M)$  also drawn at random from the same population as the training data are given.

Show that

$$\mathrm{E} \left[ \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{x}_i^\top \hat{\beta})^2 \right] \leq \mathrm{E} \left[ \frac{1}{M} \sum_{i=1}^M (\tilde{y}_i - \tilde{\mathbf{x}}_i^\top \hat{\beta})^2 \right],$$

where  $\hat{\beta}$  is the ordinary least squares estimate obtained for the training data assuming a linear regression model with  $p$  regression coefficients and the expectations are over all that is random in each expression.

### Sub-task 3:

For a regression model with  $N$  observations, the coefficient of determination  $R^2$  is defined as

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2},$$

where  $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$  and  $\hat{y}_i$  is the predicted value for the  $i$ th observation.

Show that the coefficient of determination  $R^2$  is equal to the square of the correlation between  $X$  and  $Y$  in the simple linear regression case, where

$$Y = \beta_0 + \beta_1 X + \epsilon$$

and the regression coefficients  $(\beta_0, \beta_1)^\top$  are estimated using ordinary least squares.

### Sub-task 4:

Assume a linear regression model with

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

where  $\mathbf{y} \in \mathbb{R}^N$ ,  $\mathbf{X} \in \mathbb{R}^{N \times p}$ ,  $\beta \in \mathbb{R}^p$ ,  $\epsilon \in \mathbb{R}^N$  and

1.  $E(\epsilon) = \mathbf{0}$
2.  $\text{Var}(\epsilon) = \sigma^2 \mathbf{I}$
3.  $\mathbf{X}$  deterministic with full column rank.

Show that in the case the scaled design matrix  $\frac{1}{\sqrt{N}}\mathbf{X}$  is orthonormal (i.e.,  $\frac{1}{N}\mathbf{X}^\top \mathbf{X} = \mathbf{I}$ ) and  $\beta \neq 0$ , that there always exists a value of the penalty parameter  $\lambda$  such that the ridge estimator

$$\hat{\beta}^{\text{ridge}}(\lambda) = \left( \frac{1}{N} \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \frac{1}{N} \mathbf{X}^\top \mathbf{y}$$

has lower mean squared error than the OLS estimator.

Note that the mean squared error is given by:

$$E[(\tilde{\beta} - \beta)^\top (\tilde{\beta} - \beta)]$$

with  $\tilde{\beta}$  either the OLS estimator or the ridge estimator for a given  $\lambda$ .

### Sub-task 5:

Assume the following data generating process with  $X$  drawn from some distribution and

$$\begin{aligned} Y|X &\sim f(X) + \epsilon, \\ \epsilon &\sim N(0, \sigma^2). \end{aligned}$$

A sample of  $N$  pairs  $(x_i, y_i)$  independently drawn from the data generating process are given.

Construct an estimator for  $f$  linear in  $y_i$ ,

$$\hat{f}(x_0) = \sum_{i=1}^N \ell_i(x_0, \mathcal{X}) y_i,$$

where the weights  $\ell_i(x_0, \mathcal{X})$  do not depend on  $y_i$ , but depend on the entire training sequence  $x_i$ , denoted here by  $\mathcal{X}$ .

- Show that  $k$ -nearest-neighbor regression is a member of this class of estimators and explicitly describe the weights  $\ell_i(x_0, \mathcal{X})$ .
- Show that linear regression given by

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

is a member of this class of estimators and explicitly describe the weights  $\ell_i(x_0, \mathcal{X})$ .

### Sub-task 6:

Assume that

$$Y = f(X) + \epsilon, \quad f(X) = X_1$$

with  $\epsilon \sim N(0, \sigma^2)$  and  $X$  is independent and uniformly distributed on  $[-1, 1]^p$ .

$X_1$  denotes the first element of the  $p$ -dimensional vector  $X$ .

- Use simulations to approximate the expected prediction error (EPE) given by

$$\begin{aligned}\text{EPE}_{\hat{f}_T}(x_0) &= \mathbb{E}[(Y - \hat{f}_T(x_0))^2 | X = x_0] \\ &= \text{Var}(Y|X = x_0) + \mathbb{E}_T[(\mathbb{E}(Y|X = x_0) - \hat{f}_T(x_0))^2 | X = x_0] \\ &= \text{Var}(Y|X = x_0) + [\text{Bias}_T^2(\hat{f}_T(x_0)) + \text{Var}_T(\hat{f}_T(x_0))]\end{aligned}$$

with  $\mathcal{T}$  the training sample at  $x_0 = \mathbf{0}$  based on  $m = 1000$  repetitions

- for a training sample size of  $N = 100$ ,
- the number of dimensions  $p$  varying from 1 to 10,
- with a standard deviation  $\sigma$  of either zero or one,
- using linear models (with an intercept) as well as 1-nearest neighbors to estimate  $f(X)$ .
- Assess and interpret the impact of the dimension on the EPE in dependence of the method used as well as the value of the standard deviation.

#### Sub-task 7:

Use the diabetes data set to fit different linear models. The data set is available in the R package **lars** and can be loaded using:

```
> data("diabetes", package = "lars")
```

The dependent variable is contained in `diabetes$y`, the model matrix in `diabetes$x` for the linear regression.

- Set a random seed and split the data set into a training and test data set such that 300 observations are used for training and the remaining ones for testing. Explain why it might be good to randomly select 300 observations from the available data set instead of using the first 300.
- The covariates are only available in standardized form. Explain if this is an issue for the subsequent analysis.
- Analyze the pairwise correlation structure between the covariates as well as the covariates and the dependent variable. Interpret the results and explain how these correlations impact model selection.
- Fit a linear regression model containing all explanatory variables. Inspect the model and evaluate the in-sample fit as well as the performance on the test data based on the mean squared error (MSE).
- Fit a smaller model where only the covariates are contained which according to a  $t$ -test are significant at the 5% significance level conditional on all other variables being included. Evaluate the performance in-sample as well as on the test data. Compare this model to the full model using an  $F$ -test. data and compare this model to the full model using an  $F$ -test.

#### Sub-task 8:

Use again the diabetes data set available in the R package **lars** as in the previous sub-task. Split again the data set into a training and test data set such that 300 observations are used for training and the remaining ones for testing after setting a random seed.

- Use backward-stepwise regression based on the AIC to select a suitable model. Evaluate the performance in-sample as well as on the test data and compare this model to the full model using an  $F$ -test.
- Use best subset selection to select a suitable model based on the AIC. Evaluate the performance in-sample as well as on the test data and compare this model to the full model using an  $F$ -test.

Summarize the results in a table containing the regression coefficients of the full model as well as the model selected with stepwise regression and best subset selection as well as the in-sample and the test data performance and interpret the results.