Elements of Machine Learning
Exercise Sheet 1
Winter Term 2023/2024

William LaCroix - wila00001@stud.uni-saarland.de - 7038732
Philipp Hawlitschek - phha00002@stud.unisaarland.de - 7043167

Problem 2 (T, 8 points). Error Measures.

In the regression setting, we most commonly use RSS as an error measure. Consider instead the following loss function L,

$$L\left(\beta, r\right) = \frac{1}{n}\sum_{i}^{n} r_i \left(y_i - \beta_0 - x_i\beta\right)^2 \tag{2.1}$$

for a target y and single predictor $X \in \mathbb{R}^n$.

1. [1pts] What is the effect of the parameters r_i?

r_i has the effect of weighting individual data points, punishing some more than others in terms of contribution to the loss function.

2. [3pts] Derive the minimizer $\hat{\beta}$ of Eq.(2.1) when we keep r fixed.

$$\frac{\delta L}{\delta \beta_0} = -\frac{2}{n}\sum_{i}^{n} r\left(y_i - \beta_0 - x_i\beta\right)$$

$$0 = -\frac{2r}{n}\sum_{i}^{n} y_i + \frac{2r}{n}\sum_{i}^{n}\beta_0 + \frac{2r}{n}\sum_{i}^{n} x_i\beta$$

$$\frac{2r}{n}\sum_{i}^{n}\beta_0 = \frac{2r}{n}\sum_{i}^{n} y_i - \frac{2r}{n}\sum_{i}^{n} x_i\beta$$

$$2r\beta_0 = 2r\bar{y} - 2r\beta\bar{x}$$

$$H1: \quad \hat{\beta_0} = \bar{y} - \beta\bar{x}$$

$$L = \frac{1}{n}\sum_{i}^{n} r\left(y_i - \bar{y} + \beta\bar{x} - \beta x_i\right)^2$$

$$= \frac{1}{n}\sum_{i}^{n} r\left(y_i - \bar{y} - \beta\left(x_i - \bar{x}\right)\right)^2$$

$$\frac{\delta L}{\delta \beta} = -\frac{2r}{n}\sum_{i}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y} - \beta\left(x_i - \bar{x}\right)\right)$$

$$0 = -\frac{2r}{n}\sum_{i}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) + \frac{2r}{n}\sum_{i}^{n}\beta\left(x_i - \bar{x}\right)^2$$

$$\frac{2r}{n}\sum_{i}^{n}\beta\left(x_i - \bar{x}\right)^2 = \frac{2r}{n}\sum_{i}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

$$H2: \quad \hat{\beta} = \frac{\sum_{i}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i}^{n}\left(x_i - \bar{x}\right)^2}$$

3. [3pts] Consider the following choices for r. Explain the effect of each choice, as well as what purpose it could serve.

- We set each r_i to an integer value $r_i \in \mathbb{Z}$ with $r_i > 1$.
  If for example we were performing regression on timeseries data, and wanted more recent data points to be more relevant to the regression function, we could have r_i increase as $i \to n$ in order to more heavily punish the error of the points we want to prioritize.
- Assuming that we know the noise variance σ_i of each data sample x_i, we set $r_i = \frac{1}{\sigma_i^2}$.
  If we can divide the sample residual by the square of its variance, then we are weighting the mean square error estimate by increasing the effect on the loss function of data points with smaller variance.

- Assume that we have two different datasets X_1 and X_2. From X_1, we estimate the probability density function over X as p, and from X_2 we estimate the density as q. We set $r_i = \frac{p(x_i)}{q(x_i)}$
‘ Here we are trying to see how well our estimation of X fits either dataset. If both fit the data equally well, then $r = 1$, and there is no change to the loss. For a given data point, if $p > q$, then X_1 is a better fit, and we weight x_i more highly. If $p < q$, then X_2 is a better fit, and we weight x_i lower.

4. [1pts] List three main assumptions that we rely upon in ordinary linear regression. Is there an assumption that we can address by using L instead of RSS?

We assume:

1. there exists a linear relationship between variables and output
2. errors are homoscedastic and uncorrelated
3. errors share a common variance term

As in 3.3 above, if we know the noise (σ_i) of each data point in our sample, we can account for differences in variance across the sample.