Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

CSSH Computational Social Sciences and Humanities | RWTH AACHEN UNIVERSITY

Exercise 6

Social Data Science

1 Logistic Regression

Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergraduate GPA, and $Y =$ receive an A. We fit a logistic regression and produce the following estimated coefficients $\hat{\beta}_0 = -6$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$.

a) Estimate the probability that a student who studies for 40 h and has an undergraduate GPA of 3.5 gets an A in the class.

b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

2 Gradient Descent in Logistic Regression

Recall that in logistic regression, the gradient descent optimization aims to maximize the log-likelihood

$$\log p(y|x) = y \log \hat{y} + (1 - y) \log(1 - \hat{y}),$$

where $\hat{y} = \sigma(\beta x + \beta_0)$. Derive the update rule for the gradient descent that optimizes this log-likelihood!

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

3 Bias vs Variance

Suppose you are given a dataset of $n = 100$ observations, containing a single predictor x and a quantitative response y, which you have split into a traing and test set. Further assume you fit a simple linear regression model $y \simeq \beta_0 + \beta_1 x$ to the training data, as well as a separate cubic regression $y \simeq \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$.

a) Suppose that the true relationship between x and y is linear. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer!

b) Answer a) using test rather than training RSS.

c) Suppose that the true relationship between x and y is not linear but polynomial, though we don't know to which polynomial degree. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

d) Answer c) using test rather than training RSS.