University of Stuttgart | ANALYTIC COMPUTING

Machine Learning (SS 23)

Assignment 5: Linear Methods for Classification

Mojtaba Nayyeri
Mojtaba.Nayyeri@ipvs.uni-stuttgart.de

Akram Sadat Hosseini
Akram.Hosseini@ipvs.uni-stuttgart.de

Nadeen Fathallah
Nadeen.Fathallah@ipvs.uni-stuttgart.de

Rodrigo Lopez Portillo Alcocer
rodrigo.lopez-portillo-alcocer@ipvs.uni-stuttgart.de

Tim Schneider
timphillip.schneider@ipvs.uni-stuttgart.de

Osama Mohammed
osama.mohammed@ipvs.uni-stuttgart.de

Daniel Frank
daniel.frank@ipvs.uni-stuttgart.de

Make sure to list the full names of all participants, matriculation number, study program, and B.Sc. or M.Sc on the first page. Optionally, you can *additionally* upload source files (e.g., PPTX files). If you have any questions, feel free to ask them in the exercise forum in ILIAS.

Submission is open until Monday, 5th of June 2023, 12:00 noon.

Linear Regression with Regularization

We continue with the lake example from `Assignment 4`. Please follow the instructions in the notebook `05_regression_ctd.ipynb` *Task 1 (Assignment 5)*.

From Linear Regression to Classification

1. Research (e.g. see Hastie et al. chapter 4.1) and then explain what a discriminative function is and how it can be used for classification problems using the tools from linear regression.
2. In the plots on slide 28 ("Quadratic regression to the rescue") of the *Logistic Regeression* slide deck from the lecture you can see the *Pros* and *Cons* mentioned on slide 25 ("Pros and Cons"). Explain the following:
   (a) Why does masking occur for this particular dataset, even though linear decision boundaries can perfectly separate the classes?
   (b) Assume a fourth class would be lined up (plot on slide 26) in the same space. Would quadratic features be enough to avoid masking?
3. Construct a simple example for linear regression of an indicator matrix with two classes $y \in \{0, 1\}$ and calculate the optimal parameters $\hat{B}$. Which class would the optimal parameters predict on your example data ($\arg\max_{y \in \{0,1\}} \hat{Y}$)?

Log-likelihood gradient and Hessian

Consider a binary classification problem with data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$. We define

$$f(x) = \phi(x)^\top \beta \ , \quad p(x) = \sigma(f(x)) \ , \quad \sigma(z) = 1/(1 + e^{-z}).$$

$$L^{\text{nll}}(\beta) = -\sum_{i=1}^{n} \Big[y_i \log p(x_i) + (1 - y_i) \log[1 - p(x_i)] \Big]$$

where $\beta \in \mathbb{R}^d$ is a vector. (Note: $p(x)$ is a short-hand for $p(y = 1|x)$.)

Note: The gradient and Hessian are needed to compute the optimal parameters for *logistic regerssion* models. Details on how to do this will be covered in the upcoming lecture.

1. Compute the derivative $\frac{\partial}{\partial \beta} L(\beta)$. Tip: Use the fact that $\frac{\partial}{\partial z} \sigma(z) = \sigma(z)(1 - \sigma(z))$.
2. Compute the 2nd derivative $\frac{\partial^2}{\partial \beta^2} L(\beta)$.