Machine Learning and Physics
Winter Semester 2022/2023

Exercise Sheet #6

Due: 05.12.22 11h00

Roman Remme

General Regulations.

- Please hand in your solutions in groups of three people. A mix of attendees from Monday and Tuesday tutorials is fine.
- Your solutions to theoretical exercises can be either handwritten notes (scanned), or typeset using LaTeX.
- For the practical exercises, the data and a skeleton for your jupyter notebook are available at `https://github.com/hci-unihd/mlph_sheet06`. Always provide the (commented) code as well as the output, and don't forget to explain/interpret the latter. Please hand in both the notebook (`.ipynb`), as well as an exported pdf.
- Submit all your files in the Übungsgruppenverwaltung, only once for your group of three.

Linear Regression: Preliminaries

In the lecture we considered the setup

$$y = \boldsymbol{\beta}^T \mathbf{x} + \varepsilon, \tag{1}$$

with $\mathbb{E}[\varepsilon] = 0$ and $\operatorname{var}[\varepsilon] = \sigma^2$ for some data independent variance. This does not allow for a fixed offset (i.e. the model assumes $y = 0$ for $\mathbf{x} = 0$). To be more flexible, while still keeping the structure we discussed one can proceed as follows. Consider

$$y = \beta_0 + \boldsymbol{\beta}^T \mathbf{x} + \varepsilon = \tilde{\boldsymbol{\beta}}^T \tilde{\mathbf{x}} + \varepsilon, \tag{2}$$

where we have defined $\tilde{\boldsymbol{\beta}} = (\beta_0, \boldsymbol{\beta}^T)^T$ and $\tilde{\mathbf{x}} = (1, \mathbf{x}^T)^T$. Throughout these exercises (and also most often in practice) we will assume to be in this more general case, dropping the tilde from the notation.

1 Regularization and Bias

Consider a regression problem with two explanatory variables x_1, x_2, i.e. $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ and $\mathbf{x} = (1, x_0, x_1)^T$.

(a) In this setting, write down the loss function for ridge regression, penalizing the L^2-norm of $\boldsymbol{\beta}$, in components. What is the influence of the regression strength on the bias β_0? (1 pt)

(b) Oftentimes, a regularization of the bias term is unwanted. How would you modify the loss function to account for this? (1 pt)

(c) Which shapes in $\mathbb{R}^3$ do the regularization contours (i.e. sets of parameters with equal regularization penalty) of versions (a) and (b) have? (1 pt)

2 Estimating Parameter Relevance

There exist many approaches to estimate the importance of the individual features after learning the parameters (e.g. hypothesis tests, close inspections of the posteriors in a Bayesian setup, ...). Here you will be implementing an approach that follows from a direct intuitive motivation. In the permutation test to test whether the i-th feature is relevant, you randomly permute it among the data points. Starting from $\mathbf{X} \in \mathbb{R}^{p \times N}$, you then get $\tilde{\mathbf{X}} \in \mathbb{R}^{p \times N}$ with $\tilde{\mathbf{X}}_{i,:} = \pi(\mathbf{X}_{i,:})$, i.e. the i-th row of the data matrix is permuted (where $\pi(\cdot)$ indicates the permutation operation).

You will apply this to `vostok.txt`, data measured from air trappend in ice-cores extracted in Antarctica. In it, we attempt to estimate the global temperature anomaly given three features: The age of the air, the carbon dioxide concentration and the dust concentration.

For each of the p features, create a data matrix $\tilde{\mathbf{X}}$ with one of the rows permuted. Fit a new linear regression and compare it to the sum-of-squared residuals of the original $\mathbf{X}$. If they remain roughly equal then this suggests that the feature is irrelevant. Which feature is most important and which is least relevant? (3 pts)

3 σ^2 Estimation and Heteroscedastic Noise

(a) Maximum Likelihood. Focusing on a single point $(y_n, \mathbf{x}_n)$ our linear regression model simplifies to

$$y_n = \boldsymbol{\beta}^T \mathbf{x}_n + \varepsilon_n. \tag{3}$$

If we assume that $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$, this is equivalent to the assumption that $y_n \sim \mathcal{N}(\boldsymbol{\beta}^T \mathbf{x}_n, \sigma^2)$. The logarithm of $p(y_n|\boldsymbol{\beta}, \sigma^2)$ is known as the log-likelihood. Having observed N data points this formulation generalizes to a sum of log-likelihoods and we can learn $\boldsymbol{\beta}$ by maximizing the logarithm of

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \sum_{n=1}^{N} \log \mathcal{N}\left(y_n|\boldsymbol{\beta}^T \mathbf{x}, \sigma^2\right). \tag{4}$$

Show that we are solving the same objective as in the SSQ formulation (just with a different scaling factor) and get the same solution for $\boldsymbol{\beta}$. This formulation is known as the *Maximum Likelihood* approach, as we learn the parameters that maximize the likelihood of the data. (2 pts)

(b) Estimation of σ^2. Estimating σ^2 then analogously consists of finding the $\hat{\sigma}^2$ that maximizes this log-likelihood given the estimates $\hat{\boldsymbol{\beta}}$, i.e.

$$\hat{\sigma}^2 = \arg\max_{\sigma^2} \sum_{n=1}^{N} \log \mathcal{N}\left(y_n|\hat{\boldsymbol{\beta}}^T \mathbf{x}, \sigma^2\right). \tag{5}$$

Solve this and relate the result to the SSQ residual formulation from the lecture.

(2 pts)

(c) Bonus: Heteroscedastic Noise. The standard formulation of linear regression is of homoscedastic noise, i.e. the variances of the observation noise is independent of $\mathbf{x}$. A generalization is to have a data point dependent variance on the observation noise, i.e. we have

$$y_n = \boldsymbol{\beta}^T \mathbf{x}_n + \varepsilon_n, \tag{6}$$

with $\mathbb{E}[\varepsilon_n] = 0$ and $\mathrm{var}[\varepsilon_n] = \sigma_n^2$, which is known as *heteroscedastic noise*. Give the sum-of-squares problem in that case and derive mean and covariance structure of the $\hat{\boldsymbol{\beta}}$ in that case. (3 pts)

4 Visualize Regularization Contours (10 pt)

For two dimensional parameter vectors β we can visualize the error/loss surface of linear regression using contour plots. In this exercise you will create a set of such plots in order to familiarize yourself further with the influence of regularization. You can visualize the contours for example via `plt.contour` or `plt.contourf`.[1]

[1] See `https://matplotlib.org/stable/gallery/images_contours_and_fields/contour_demo.html` for an example.

(a) Plot the Ridge regression regularization term as well as the Lasso[2] regularization term for $\beta_1, \beta_2 \in [-1, 3]$. (2 pts)

(b) For the data set `linreg.npz` plot the sum of squares (SSQ) of a linear regression as a function of $\boldsymbol{\beta}$ over the same range as in **i)**, i.e. over the grid $[-1, 3] \times [-1, 3]$. (2 pts)

(c) Plot the ridge and Lasso loss functions, i.e. $\mathrm{SSQ}(\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_2^2$ and $\mathrm{SSQ}(\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_1$, for $\lambda \in \{0, 10, 50, 100, 200, 300\}$ in the same $\boldsymbol{\beta}$ grid as before and *discuss your observations!* (2 pts)

5 CT Reconstruction

One application of linear regression is the reconstruction of CT-Scans. In this task, you will do this on simulated data in the 2D case. You are given a sinogram $Y \in \mathbb{R}^{ar}$, a matrix where each row corresponds to a (1D) projection of the image consisting of r detector readouts along one of a distinct, evenly spaced angles. Additionally, you are given the design matrix $\mathbf{X} \in \mathbb{R}^{p \times ar}$. Excluding noise, one has $Y = I\mathbf{X}$, with the image $I \in \mathbb{R}^p$ which should be reconstructed.

(a) What is the interpretation of a column of $\mathbf{X}$? Visualize a choice of four columns as images. (1 pt)

(b) Solve the reconstruction problem with linear regression without any regularization and with ridge regression. What do you observe?
(3 pts)

6 Bonus: X-Ray Free-Electron Lasers

Imagine the reconstruction problem from task 5, but without the knowledge about which detections correspond to which orientations of the sample. This scenario actually happens in the analysis of X-ray free-electron laser data: The laser is aimed at a sample (a tiny protein crystal), and the diffraction pattern of a short high-energy pulse is recorded. This is repeated many times, but never is the orientation of the sample known. Nevertheless, it's possible to reconstruct the structure of proteins from such data[3].

How would you approach this problem? Try to reconstruct the image from the data of task 5, after shuffling the sinogram along the angle axis. (5 pts)

[2] The abbreviation comes from *least absolute shrinkage and selection operator*.

[3] There are more complications in the analysis of the from diffraction patterns, such as solving the phase problem. Here, we merely focus on the reshuffling problem, reusing the data from the previous exercise.