Assignment 3

Statistical Machine Learning

Sebastian Bordt / David Künstle / Martina Contisciani / Nicolò Ruggeri / Rabanus Derr / Prof. Ulrike von Luxburg

Summer term 2020 — due on **May 14th at 14:00**

Exercise 1 (Linear Regression, 1 + 1 + 0.5 + 0.5 + 0.5 + 1.5 +2 = 7 points)
Let $X \in \mathbb{R}^{n \times d}$ be a data matrix. In this exercise, we are going to prove that the least-squares optimization problem

$$\min_{w \in \mathbb{R}^d} \| Y - Xw \|_2^2 \tag{1}$$

is convex (compare Proposition 5 in the lecture slides). To warm up, we recap multidimensional derivatives.

(a) Let $f(X) = a^T X$, where $X \in \mathbb{R}^3$ is a column vector, and $a^T = [2, -1, 5]$. Compute

$$\frac{\partial f}{\partial X}$$

(b) Let $f(X) = X^T A X$, where $X \in \mathbb{R}^2$ is a column vector of two elements and

$$A = \begin{pmatrix} 1 & 2 \\ 5 & 3 \end{pmatrix}$$

compute

$$\frac{\partial f}{\partial X}$$

(c) Argue that the optimization problem (1) is defined over a convex domain.

(d) Does the optimization problem (1) have any constraints? If not, does this mean that this is an unconstrained optimization problem?

(e) Is the objective function in (1) differentiable? If yes, how many times?

(f) Show that $X^T X$ is positive semi-definite and that $\text{rank}(X^T X) = \text{rank}(X)$.

(g) Show that the optimization problem (1) is convex.
Hint: What is left to show? Use an appropriate criterion.

Exercise 2 (Linear, Ridge, and Lasso in Python, 1+2+2+2+3+1+2 = 13 points)

In this exercise you will implement linear regression, ridge regression and the lasso.

(a) Let

$$x \sim \text{Unif}([0, 2))$$
$$y(x) \sim 2 \sin 2x + 0.1 * \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, 2)$. In the notebook you will find the code that samples 100 points from this distribution and saves them in (`xs`, `ys`). Do a scatter plot of the sampled points.

(b) Write two functions

```
def ridge_regression(xs, ys, lam=1): ...
```

and

```
def lasso_regression(xs, ys, lam=1): ...
```

that, given an $n \times D$ matrix `xs` and an $n \times 1$ vector `ys`, return the ridge regression resp. lasso weights ω as a $D \times 1$ vector.

(c) For $\lambda \in \{0.1, 1, 10\}$, compute the ridge regression for $(\mathtt{xs}, \mathtt{ys})$ as in a) and compare the mean squared error (MSE). In three different plots, plot the predictions for the different λ. The plot should contain the scatter plot of the points and the predicted line. Repeat the MSE calculations and the plotting for lasso regression.

(d) Compute the weights for all datasets in the provided list using linear, ridge ($\lambda = 1$) and lasso ($\lambda = 1$) regression. Scatter plot the absolute weight vectors, such that the weight dimensions correspond to the x- and y-axis. Use logarithmic axis scale and appropriate limits.

(e) For `xs` as in a) compute the representation of `xs` in the basis given by

$$\left\{1, \mathtt{xs}, \mathtt{xs}^2\right\}$$

and then perform ridge regression on this basis. Use $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$ and compare the MSE. Plot the best prediction. Set $\lambda = 0$, to see what happens with the linear regression.

(f) In `xs_test` and `ys_test` you will find 20 new samples. Using the same base as in (e), train linear and ridge regression for $\lambda \in \{0.001, 0.01, 0.1\}$ on the original samples and then predict on these new test points. Which method works best?

(g) Let

$$x \sim \mathrm{Unif}([0, 2)^2)$$
$$y(x) \sim 2x_1^2 + 2x_2 + 1 + 0.1 * \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, 1)$. For every x compute the representation of x in the basis given by

$$\left\{1, x_1, x_1^2, x_2, x_2^2, x_1 x_2\right\}$$

For $\lambda \in \{0.001, 0.01, 0.1, 1, 10\}$ perform ridge regression on this basis, using x and y as inputs, and plot the MSE as a function of λ.