Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNIA Winter Term 2018/2019

Exercise Sheet 5

Machine Learning Basics - MLE, Cross-Validation, Logistic Regression

Deadline: 03.12.2018, 23:59

Exercise 5.1 - Maximum Likelihood Estimation (MLE) (1.5 + 1.5 + 2 = 5 points)

a) Consider the density function of a **univariate Gaussian distribution**

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

where μ is the *mean* and σ^2 is the *variance*.

Let's say you're given N samples (i.e. $x_1, x_2, x_3, ..., x_N$) which are drawn from the above stated distribution. Also, you can assume that these samples are **i.i.d** (i.e. independent and identically distributed). Now, please derive the *MLE step-by-step* for:

i) *mean* (μ)

ii) *variance* (σ^2)

b) Consider the linear regression problem from the lecture. The ground truth y_i can be considered as being generated by $y_i = \mathbf{w}_{gt}^T \mathbf{x}_i + \epsilon_i$, where ϵ_i is a random noise which follows a standard Gaussian distribution and $\mathbf{w}_{gt}$ denotes the ground truth weight. Prove that MLE of $\mathbf{w}$ with $p(y_i|\mathbf{x}_i)$ is equivalent to minimizing the MSE for $\mathbf{w}$ from the lecture.

Exercise 5.2 - Validation & Cross-Validation (2 + (1 + 1) = 4 points)

For the task of Multiple Linear Regression in Project-1, we split our dataset and allocated a portion of it for validation purposes. Such a process of setting aside a portion of the dataset is known as *holdout method* and the dataset is called holdout dataset.

a) Why would one need cross-validation instead of the *holdout method*? Explain two scenarios where cross-validation is needed.

b) Assume that we want to find an optimal capacity of our model for the task of linear regression, with possible choices for the order of the polynomial as: {1, 5, 9}.

Now, assume that we want to do this hyperparameter selection using *k-fold cross-validation* (with $k = 5$), instead of the *holdout method*. Given this setting, pictorially explain the steps that are involved in this 5-fold cross-validation, along with a brief explanation. Also, explain how would you compute a single final score (e.g. MSE) for each of the hyperparameters (i.e. order of the polynomial), so that we can compare the performance of these models and choose the best one.

Exercise 5.3 - Logistic Regression (1 points)

In chapter 5, we briefly learned about *Logistic Regression*. It uses the so-called *Logistic function* (https://en.wikipedia.org/wiki/Logistic_function) instead of a polynomial.

a) Why would one need or choose (multinomial) logistic regression over linear regression irrespective of the usage of regularization? Briefly explain.

Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your exercise.

- You **have to** submit a solution of this assignment sheet as a team of 2-3 members.
- Hand in a **single** PDF file with your solutions to the tasks.
- Therefore, make sure to write the name and matriculation ID of each of the members in your team.
- The solution must be uploaded by only **one** of your team members to the course website.
- If you have any trouble with the submission, contact your tutor **before** the deadline.

Plagiarism of any form is not tolerated. If you refer something from the web, you must give proper credit by citing the source. Lack of this would be considered plagiarism. In such a case, the whole sheet would be awarded zero points and a warning is given. If this act is repeated again, then the whole team is excluded from the course.