Machine Learning 2019-2020

Home Assignment 2

Yevgeny Seldin   Christian Igel
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **3 December 2019 22:00**. You must submit your *individual* solution electronically via the Absalon home page.

A solution consists of:

- A PDF file with detailed answers to the questions, which may include graphs and tables if needed. Do *not* include your source code in the PDF file.

- A .zip file with all your solution source code with comments about the major steps involved in each question (see below). Source code must be submitted in the original file format, not as PDF.

- IMPORTANT: Do NOT zip the PDF file, since zipped files cannot be opened in the speed grader. Zipped pdf submissions will not be graded.

- Your PDF report should be self-sufficient. I.e., it should be possible to grade it without opening the .zip file. We do not guarantee opening the .zip file when grading.

- Your code should be structured such that there is one main file (or one main file per question) that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.

- We strongly recommend that your code is written in python 3. If you prefer to use another mainstream language (e.g. R, Matlab, C/C++), you can do so at your own risk and responsibility.

- Your code should include a README text file describing how to compile and run your program, as well as a list of all relevant libraries needed for compiling or using your code.

- Handwritten solutions will not be accepted, please use the provided latex template to write your report.

1 Illustration of Hoeffding's Inequality (10 points)

Go back to Question 2 in Home Assignment 1 and produce the figure you had there.

1. In the same figure plot the Hoeffding's bound[1] on $\mathbb{P}\big(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\big)$ for the same values of α.

2. Compare Hoeffding's bound with the other three plots.

3. For $\alpha = 1$ and $\alpha = 0.95$ compare the exact probability $\mathbb{P}\big(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\big)$ you have calculated in Home Assignment 1 with the Hoeffding's bound. (No need to add this one to the plot.)

Do not forget to put axis labels and a legend in your plot!

2 The effect of scale (range) and normalization of random variables in Hoeffding's inequality (10 points)

Prove that Corollary 2.5 in Yevgeny's lecture notes (simplified Hoeffding's inequality for random variables in the $[0, 1]$ interval) follows from Theorem 2.3 (general Hoeffding's inequality). [Showing this for one of the two inequalities is sufficient.]

3 Distribution of Student's Grades (15 points)

A student submits 7 assignments graded on the 0-100 scale. We assume that each assignment is an independent sample of his/her knowledge of the material and all scores are sampled from the same distribution. Let $X_1, \ldots, X_7$ denote the scores and $\hat{Z} = \frac{1}{7}\sum_{i=1}^{7} X_i$ their average. Let p denote the unknown expected score, so that $\mathbb{E}[X_i] = p$ for all i. What is the maximal value z, such that the probability of observing $\hat{Z} \leq z$ when $p = 50$ is at most $\delta = 0.05$?

1. Use Markov's inequality to answer the question. (Hint: in order to get a lower bound you have to consider the random variable $\hat{Q} = 100 - \hat{Z}$.)

---

[1]Hoeffding's bound is the right hand side of Hoeffding's inequality.

2. Use Chebyshev's inequality to answer the question. (You can use the fact that the variance of a random variable $X \in [a, b]$ is maximized when $X = a$ with probability $1/2$ and $X = b$ with probability $1/2$. In other words, let Y be a random variable, such that $\mathbb{P}(Y = a) = 1/2$ and $\mathbb{P}(Y = b) = 1/2$, then for any random variable $X \in [a, b]$ we have $\text{Var}[X] \leq \text{Var}[Y]$.)

3. Use Hoeffding's inequality to answer the question.

4. Which of the three inequalities provide a non-vacuous value of z? (You know without any calculations that for any $z < 0$ we have $\mathbb{P}(Z \leq z) = 0$, so any bound smaller than 0 is useless.)

4 The Airline Question (15 points)

1. An airline knows that any person making a reservation on a flight will not show up with a probability of 0.05 (5 percent). They introduce a policy to sell 100 tickets for a flight that can hold only 99 passengers. Bound the probability that the number of people that show up for a flight will be larger than the number of seats (assuming they show up independently).

2. An airline has collected an i.i.d. sample of 10000 flight reservations and figured out that in this sample 5 percent of passengers who made a reservation did not show up for the flight. They introduce a policy to sell 100 tickets for a flight that can hold only 99 passengers. Bound the probability of observing such sample and getting a flight overbooked.

   Hint: there are multiple ways to approach this question. We will guide you through one option. Put attention that the true probability, let's call it p, of showing up for a flight is unknown. We consider two events: the first is that in the sample of 10000 passengers, where each passenger shows up with probability p, we observe 95% of show-ups. The second event is that in the sample of 100 passengers, where each passenger shows up with probability p, everybody shows up. Note that these two events are independent. Bound the probability that they happen simultaneously assuming that p is known. And then find the worst case p (you can do this numerically). With a simple approach you can get a bound of around 0.61. If you are careful and use the right bounds you can get down to around 0.0068.

   It is advised to visualize the problem (the $[0, 1]$ interval with 0.95 point for the 95% show-ups and 1 for the 100% show-ups and p somewhere in $[0, 1]$). This should help you to understand the problem; to understand where the worst case p should be; and to understand in what direction of inequalities you need.

5 Logistic Regression

5.1 Cross-entropy error measure (12 points)

Read section 3.3 in the course textbook (Abu-Mostafa et al., 2012). You can also find a scanned version of the chapter on Absalon. Solve exercise 3.6 on page 92 in the course textbook. The *in-sample error* E_{in} corresponds to what we call the empirical risk (or training error).

5.2 Logistic regression loss gradient (14 points)

Solve exercise 3.7 on page 92 in the course textbook (Abu-Mostafa et al., 2012) .

The book assumes labels in $\{-1, 1\}$. Solve exercise 3.7 again assuming the labels $\{0, 1\}$, which leads to

$$\nabla E_{\text{in}}(\boldsymbol{w}) = -\frac{1}{N} \sum_{n=1}^{N} \left[y_n - \theta(\boldsymbol{w}^{\text{T}} \boldsymbol{x}) \right] \boldsymbol{x}_n \ .$$

Hints: Do not forget the "Argue . . . one." part in the exercise for both parts of this question. For the $\{0, 1\}$ case the slides provide the answer, you just need to add an explanation and intermediate steps.

5.3 Logistic regression implementation (12 points)

Implement logistic regression as presented in the lectures, see also Example 3.3 on page 95 in Abu-Mostafa et al. (2012). Attach the source code and briefly describe your implementation in the report. You may present the most important parts of your implementation in the report. The answer to this question may not exceed one page in the report.

5.4 Iris flower data (12 points)

Please download the data sets `IrisTrainML.dt` and `IrisTestML.dt` from the course homepage. These data sets have been generated from the famous Iris flower data set Anderson (1936), which has been used as an example for classification algorithms since the work of Ronald Fisher Fisher (1936). However, instead of the original four input features only two are considered. Furthermore, a feature has been rescaled and some examples have been removed.

The data set describes three different species of Iris, namely Iris setosa, Iris virginica and Iris versicolor. That is, we have a three class classification problem.

Each line in the data files corresponds to one flower. The first two columns of our version of the data are the lengths and the widths of the sepals. Sepals are modified leaves that are part of the calyx of a flower. In our modified version of the data set, the length is measured in millimeters and the width in centimeters. The last column encodes the species.

Figure 1: An example of an Iris versicolor taken from Wikipedia.

Apply logistic regression to the Iris flower data. Remove class 2 from the training and test data sets. The resulting data sets should contain 62 and 26 patterns, respectively. That is, you are supposed to solve a binary classification task.

Report the training and test error as measured by the 0-1 loss. Furthermore, report the three parameters of the (affine) linear model.

References

Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning from Data.* AMLbook, 2012.

E. Anderson. The species problem in iris. *Annals of the Missouri Botanical Garden*, 23(3):457–509, 1936.

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.