Machine Learning 2017/2018

Home Assignment 2

Yevgeny Seldin, Christian Igel
Department of Computer Science, University of Copenhagen

The deadline for this assignment is **5 December 2017**. 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 (Matlab / R / Python scripts / C / C++ / Java / etc.) 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.
- 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.

1 Illustration of Hoeffding's Inequality

1. Make 1,000,000 repetitions of the experiment of drawing 20 i.i.d. Bernoulli random variables $X_1, \ldots, X_{20}$ (20 coins) with bias $\frac{1}{2}$.

2. Plot the empirical frequency of observing $\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha$ for $\alpha \in (0.5, 0.55, 0.6, \ldots, 0.95, 1)$.

3. Explain why the above granularity of α is sufficient. I.e., why, for example, taking $\alpha = 0.51$ will not provide any extra information about the experiment.

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

5. In the same figure plot the Markov's bound[2] on $\mathbb{P}\{\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\}$.

6. Compare the three plots.

7. For $\alpha = 1$ and $\alpha = 0.95$ calculate the exact probability $\mathbb{P}\{\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\}$ and compare it 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

Prove that Corollary 2.4 in Yevgeny's lecture notes (simplified Hoeffding's inequality for random variables in the $[0, 1]$ interval) follows from Theorem 2.2 (general Hoeffding's inequality).

3 Probability in Practice

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

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

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, lets 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.

4 Logistic regression

4.1 Cross-entropy error measure

Read section 3.3 in the course textbook [1] (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).

4.2 Logistic regression loss gradient

Solve exercise 3.7 on page 92 in the course textbook.

4.3 Logistic regression implementation

Implement logistic regression as presented in the lectures, see also Example 3.3 on page 95 in [1]. 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.

4.4 Iris flower data

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 [2], which has been used as an example for classification algorithms

since the work of Ronald Fisher [3]. 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 an test error as measured by the 0-1 loss. Furthermore, report the three parameters of the (affine) linear model.

References

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

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

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