Machine Learning A
2022-2023
Home Assignment 3

Sadegh Talebi   **Christian Igel**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **27 September 2022, 18:00**. 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 full source code in the PDF file, only selected lines if you are asked to do so.

- 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. The programming language of the course is Python.

- IMPORTANT: Do NOT zip the PDF file, since zipped files cannot be opened in 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.

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

1 Use of Generalization Bounds to Split a Sample into Training and Test Sets (20 points)

In this question you will analyze one possible approach to the question of how to split a dataset S into training and test sets, S^{train} and S^{test}. As we have already discussed, overly small test sets lead to unreliable loss estimates, whereas overly large test sets leave too little data for training, thus producing poor prediction models. The optimal trade-off depends on the data and the prediction model.

Assume that you have a fixed split of S into S^{train} and S^{test}, where the size of S^{test} is n^{test}. You train a model $\hat{h}^*_{S^{\text{train}}}$ on S^{train} using whatever procedure you want. Then you compute the test loss $\hat{L}(\hat{h}^*_{S^{\text{train}}}, S^{\text{test}})$. Derive a bound on $L(\hat{h}^*_{S^{\text{train}}})$ in terms of $\hat{L}(\hat{h}^*_{S^{\text{train}}}, S^{\text{test}})$ and n^{test} that holds with probability at least $1 - \delta$.

2 Confidence Intervals for Bernoulli Distribution (15 points)

We are given a dataset `S.csv`, whose elements are drawn i.i.d. from a Bernoulli with mean μ. Let us denote this dataset by a sequence $S = (X_t)_{1 \le t \le n}$ of length n (i.e., n is the number of elements in the dataset).

1. Compute 0.97-CI (two-sided) for μ using Hoeffding's inequality. Report the numerical values of the computed intervals, and provide a brief explanation of your calculations (including any formula/procedure used).

2. [**Optional, not for hand-in**] Suppose now that we are given another dataset `U.csv`, whose elements are drawn i.i.d. from a Bernoulli with mean λ. Let us denote this dataset by the sequence $U = (Y_t)_{1 \le t \le m}$ of length m. We do not know whether U is independent of S. We are interested in determining whether $\lambda > \mu$. Can we say with probability at least 0.98 that $\lambda > \mu$? Argue why.

3 Big and Small (25 points)

Suppose you are approached by two companies, called A and B, each proposing you a classifier for a problem of interest, h_A and h_B. They also provide you the data they have used for training their classifiers. A has collected a dataset S_A of 9000 samples whereas B has collected a dataset S_B of only 1000 samples. We

assume that all the data are i.i.d. and both samples come from the same distribution. The companies provide no details on how they produced the classifiers. You have no own data to test the solutions, so instead you test h_A on S_B and h_B on S_A. You obtain $\hat{L}(h_A, S_B) = 0.03$ and $\hat{L}(h_B, S_A) = 0.06$.

1. You need to pick a classifier and provide a generalization bound that will hold with probability at least 95%. Explain which of the two classifiers you will pick and provide a generalization bound for it.

   Hint: The solution involves using a union bound.

2. Repeat the question under the requirement that the bound should hold with probability at least 99%.

4 Preprocessing (20 points)

Read section 9.1 in e-Chapter 9 of the textbook (Abu-Mostafa et al., 2012). The chapter can be downloaded from `https://absalon.instructure.com/courses/61346/files/folder/Lecture%20Notes?preview=6348356`, the login is `bookreaders` and the password the first word on page 27 of the textbook. You can also find a scanned version of the section on Absalon. It is also recommended to read Section 4.2 of the textbook (Abu-Mostafa et al., 2012) on regularization. You can also find a scanned version of the section on Absalon. Note that the *in-sample error* E_{in} corresponds to what we call the empirical risk (or training error).

4.1 9.1 (4 points)

Solve Exercise 9.1 on page w-Chap:9–1 from the textbook (Abu-Mostafa et al., 2012).

4.2 9.2 (4 points)

Solve Exercise 9.2 on page w-Chap:9–3 from the textbook (Abu-Mostafa et al., 2012).

4.3 9.4 (12 points)

Solve the first 4 parts (a)–(d) from Exercise 9.4 on page w-Chap:9–4 from the textbook (Abu-Mostafa et al., 2012).

Hints:

1. Recall that when you add normally distributed random variables their variances add up, that is, if $x_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $x_2 \sim \mathcal{N}(0, \sigma_2^2)$, then $x_1 + x_2 \sim \mathcal{N}(0, \sigma_1^2 + \sigma_2^2)$. And remember if $x \sim \mathcal{N}(0, 1)$ then $\sigma x \sim \mathcal{N}(0, \sigma^2)$.

2. You can compute the covariance in several ways. One basic way to derive it is the following. Consider the standard normally distributed random vector $\hat{\boldsymbol{x}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and its transformation $\boldsymbol{x} = \boldsymbol{A}\hat{\boldsymbol{x}}$. We have $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})$. For the exercise, find $\boldsymbol{A}$, and then the off-diagonal elements of $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ give you the covariance between x_1 and x_2.

5 Variable Importance (20 points)

Compared to many other machine learning models, a logistic model can be easily interpreted. The sign of a coefficient tells us whether the corresponding input variable has positive or negative effect on the prediction of the output class. The amount of a coefficient is related to the importance of a variable for the prediction. However, $w_i > w_j$ (using the previous notation, w_i is the i-th component of $\boldsymbol{w}$) does not simply imply that the i-th input variable is more important than the j-th input variable. Obviously, this also depends on the scaling of the input variables. For example, the importance of an input variable measuring a distance in the physical world should be independent of whether the associated unit is meters or millimeters.

The notebook `Variable importance using logistic regression.ipynb` demonstrates how to analyze the importance of the variables in a logistic regression model. Please have a close look. (Note that for a non-linear model the importances and their ranking may be different.) As argued above, we do not consider the amount of a coefficient directly, but the corresponding z-statistic, which in our case is the coefficient over its standard error. The z-statistic of a coefficient is invariant under linearly rescaling of the corresponding input variable.

In the example in the notebook, we have to deal with a categorical variable. A categorical variable takes values that correspond to a particular category (class, concept, object), for example {Orange, Apple, Banana}, and these categories are not necessarily ordered in a meaningful way. Such a variable needs to be encoded before a (generic) machine learning system processes the data. Simply encoding {Orange, Apple, Banana} by {0, 1, 2} and treating the variable as measured on an interval scale (i.e., treating the categories as numbers), does not make sense – a banana is not two times an apple.

You already heard about the most popular encoding for output categorical variables, the one-hot encoding. A one-hot encoding of $\{\text{Orange}, \text{Apple}, \text{Banana}\}$ is $\{(1,0,0)^{\mathrm{T}}, (0,1,0)^{\mathrm{T}}, (0,0,1)^{\mathrm{T}}\}$, that is $C = 3$ classes are encode by C (output) variables. In the notebook, however, $C - 1$ ("dummy") variables are used for the categorical input variable.

In this questions of the assignment, you should concisely explain why $C - 1$ variables are used instead of the one-hot encoding. Your submission should include answers to the following questions: How many solutions (i.e., optimal values for the coefficients) would the linear regression optimization problem (without regularization) have if the one-hot encoding was used? Why? Why would it be difficult to interpret the variable importance if the one-hot encoding was used?

6 [Optional, Not for Hand-in] Distribution of Student's Grades

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 $p = 60$ 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 $\hat{Q} = 100 - \hat{Z}$.)

2. Use Chebyshev's inequality to answer the question. (You can use the fact that for a random variable $X \in [a, b]$ and a random variable $Y \in \{a, b\}$ with $\mathbb{E}[X] = \mathbb{E}[Y]$ we have $\text{Var}[X] \leq \text{Var}[Y]$. In words, the variance of a random variable taking values in a bounded interval is maximized when the distribution is concentrated on the boundaries of the interval. You should determine what should be the values of $\mathbb{P}(Y = a)$ and $\mathbb{P}(Y = b)$ in order to get the right expectation and then you can obtain a bound on the variance.)

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.)

References

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