# *Machine Learning 2020-2021*

Home Assignment 1

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

The deadline for this assignment is **1 December 2020, 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. 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 Illustration of Hoeffding's Inequality (10 points)

Go back to Question 2 in Home Assignment 1 and reproduce the figures you had there.

1. In the respective figures plot the Hoeffding's bound[1] on $\mathbb{P}\left(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\right)$ for the same values of $\alpha$. (You should have one figure for bias $\mathbb{E}[X_1] = 0.5$ and another for bias $\mathbb{E}[X_1] = 0.1$.)

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

3. For $\alpha = 1$ and $\alpha = 0.95$ compare the exact probability $\mathbb{P}\left(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha\right)$ 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 $\operatorname{Var}[X] \leq \operatorname{Var}[Y]$.)

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 reservation on a flight will not show up with 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.

Attention: This is a frequentist rather than a Bayesian question. In case you are familiar with the Bayesian approach, it cannot be applied here, because we do not provide a prior on $p$. In case you are unfamiliar with the Bayesian approach, you can safely ignore this comment.

# 5 Linear classification (30 points)

## 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* $E_{\text{in}}$ 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

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

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 Log-odds (12 points)

We consider binary logistic regression. Let the input space be $\mathbb{R}^d$ and the label space be $\{0, 1\}$. Let our model $f$ with parameters $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$ model:

$$f(\mathbf{x}) = \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b) = P(Y = 1 \,|\, X = \boldsymbol{x}) \tag{1}$$

Prove that if the (affine) linear part of the model encodes the log-odds, that is, if

$$\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b = \ln \frac{P(Y = 1 \,|\, X = \boldsymbol{x})}{P(Y = 0 \,|\, X = \boldsymbol{x})} \ , \tag{2}$$

then $\sigma$ is the logistic function. That is, if $\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b$ encodes on log-scale how frequent class 1 occurs relative to class 0, then $\sigma$ is the logistic function.

### 5.4 Variable importance (12 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 $\{\text{Orange}, \text{Apple}, \text{Banana}\}$, and these categories are not necessarily ordered in a meaningful way. Such a variable needs to be encoded before a (generic) machine leanring system processes the data. Simply encoding $\{\text{Orange}, \text{Apple}, \text{Banana}\}$ 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 $\{(1, 0, 0)^\text{T}, (0, 1, 0)^\text{T}, (0, 0, 1)^\text{T}\}$, 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?

# References

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