Machine Learning A

2021-2022

Home Assignment 3

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

The deadline for this assignment is **28 September 2021, 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 Distribution of Student's Grades (25 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 $p = 60$ 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 $\mathrm{Var}[X] \leq \mathrm{Var}[Y]$. 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.)

2 How to Split a Sample into Training and Test Sets (25 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, $S^{\texttt{train}}$ and $S^{\texttt{test}}$. 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. So can we let the data speak for itself? We will give it a try.

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

2. Now we want to find a good balance between the sizes of $S^{\texttt{train}}$ and $S^{\texttt{test}}$. We consider m possible splits $\{(S_1^{\texttt{train}}, S_1^{\texttt{test}}), \ldots, (S_m^{\texttt{train}}, S_m^{\texttt{test}})\}$, where the sizes of the test sets are $n_1, \ldots, n_m$, correspondingly. For example, it could be $(10\%, 90\%), (20\%, 80\%), \ldots, (90\%, 10\%)$ splits or anything else with a reasonable coverage of the possible options. We train m prediction models $\hat{h}_1^*, \ldots, \hat{h}_m^*$, where $\hat{h}_i^*$ is trained on $S_i^{\texttt{train}}$. We calculate the test loss of the i-th model on the i-th test set $\hat{L}(\hat{h}_i^*, S_i^{\texttt{test}})$. Derive a bound on $L(\hat{h}_i^*)$ in terms of $\hat{L}(\hat{h}_i^*, S_i^{\texttt{test}})$ and n_i that holds for all $\hat{h}_i^*$ simultaneously with probability at least $1 - \delta$.

   Comment: No theorem from the lecture notes applies directly to this setting, because they all have a fixed sample size n, whereas here the sample sizes vary, $n_1, \ldots, n_m$. You have to provide a complete derivation.

3 Linear classification (50 points)

3.1 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).

3.2 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.

3.3 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 σ 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 σ is the logistic function.

3.4 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 $\{(1, 0, 0)^{\mathrm{T}}, (0, 1, 0)^{\mathrm{T}}, (0, 0, 1)^{\mathrm{T}}\}$, 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?

References

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