Machine Learning A

2023-2024

Home Assignment 3

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

The deadline for this assignment is **21 September 2023, 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 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 *SpeedGrader*. 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 Experiment design (20 points)

1. You are working at a hospital and you have collected an i.i.d. sample of 2000 patients and annotated it for presence or absence of some disease (binary annotation). You organize a competition to find a classifier for the disease. You have 20 teams that have signed up for the competition and your boss requires you to provide a confidence interval of 0.05 on the prediction accuracy of the best classifier that will hold with probability at least 95%. In other words, with probability at least 95% the estimate of the expected error should not underestimate the true expected zero-one error of the selected classifier by more than 0.05 (one-sided error). How many samples do you have to keep aside in order to satisfy this requirement, assuming that you accept 1 solution from each team? Provide a complete calculation, numerical answers without any derivations or explanations will not be accepted.

2. You have conducted the competition above, but were not satisfied with the prediction accuracy of the winner. You decided to make another competition and were very lucky to convince your boss to support annotation of another 1000 patients. You decided to release the old 2000 patients data for training and keep the new 1000 samples for evaluating the outcome of the new competition. Your boss requires from you the same confidence interval of 0.05 with probability at least 95%. How many teams can you accept to take part in the competition assuming that you accept only 1 solution from each team? Provide a complete calculation, numerical answers without any derivations or explanations will not be accepted.

2 Efficient use of the data (20 points)

So far, given a data set S, in all our theoretical results we used part of the data for training prediction rules and another part for validating them. This way some data are only used for training and some data are only used for validation. But put attention that if we would have trained a prediction rule on the validation set and validated it on the training set, we would have also gotten an unbiased estimate of the loss. (Remember: "it's not about how you call it, it's about how you use it"!). So could we use the data more efficiently?

The approach of using part of the data for training and part for validation, and then reverting the roles, somewhat resembles cross-validation. But a big word of warning would be in place here. Even though the standard cross-validation technique is widely used, it is a heuristic, and if it is used to validate too many prediction rules, it is prone to overfitting, in exactly the same way as the standard

validation technique is prone to overfitting, unless generalization bounds are used to control the overfitting.

What you will do next is inspired by cross-validation, but it is different from the standard cross-validation approach.

So, we have a data set S of size n (assume that n is even). We split the data set into two equal halves, $S = S_0 \cup S_1$. We train M models $\{h_{0,1}, \ldots, h_{0,M}\}$ on the first half of the data and validate them on the remaining half. Let $\hat{L}(h_{0,i}, S_1)$ for $i \in \{1, \ldots, M\}$ be the corresponding validation losses. Then we train another M models $\{h_{1,1}, \ldots, h_{1,M}\}$ on the second half of the data and validate them on the first half. Let $\hat{L}(h_{1,i}, S_0)$ for $i \in \{1, \ldots, M\}$ be the corresponding validation losses. Finally, we select the model $h_{j^*,i^*} = \arg \min_{j \in \{0,1\}, i \in \{1,\ldots,M\}} \hat{L}(h_{j,i}, S_{1-j})$ with the smallest validation loss.

Derive a high-probability generalization bound for the expected loss of h_{j^*,i^*}. (I.e., a bound on $L(h_{j^*,i^*})$ that holds with probability at least $1 - \delta$.)

Comment: no theorem in the lecture notes directly applies to the question, because they all assume that $\hat{L}(h, S)$ is computed on the same S for all h. You have to make a custom derivation, but it will not be very different from derivations you can find in the lecture notes.

3 How to split a sample into training and test sets (30 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.

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 $n_1, \ldots, n_m$

vary. You have to provide a complete derivation.

4 Preprocessing (30 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 (6 points)

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

4.2 9.2 (6 points)

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

4.3 9.4 (18 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 [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 $\mathrm{Var}[X] \leq \mathrm{Var}[Y]$. 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.)

References

Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from Data*. AMLbook, 2012.