Machine Learning A

2023-2024

Home Assignment 4

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

The deadline for this assignment is **28 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.

total = 55

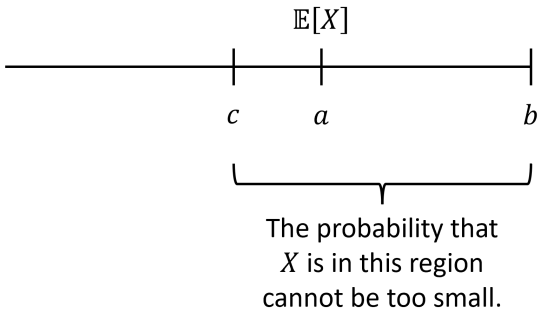

Figure 1: Illustration to the question on how to obtain a lower bound on the probability from a lower bound on the expectation.

done

1 From a lower bound on the expectation to a lower bound on the probability (5 points)

The purpose of this question is to help you understand how a lower bound on the expectation of a *bounded* random variable X can be used to derive a lower bound on the probability that X is not too small.

Let X be a random variable that is always upper bounded by b. Let $0 \leq c < a < b$. Prove that if $\mathbb{E}[X] = a$, then $\mathbb{P}(X \geq c) \geq \frac{a-c}{b}$.

Hint: see the illustration in Figure 1 and check the proof of the lower bound on the probability in Yevgeny's slides.

Informal conclusion: If we show that the expectation of a *bounded* random variable X is large, then the probability that X is large cannot be too small. (Note that if X is unbounded, the claim does not hold. You are very welcome to think why.)

Optional: With a bit more careful calculation, you can prove a tighter bound $\mathbb{P}(X \geq c) \geq \frac{a-c}{b-c}$.

Optional: You can also derive the bound using Markov's inequality. This also yields the tighter bound.

2 Learning by discretization (20 points)

We want to learn an arbitrary binary function on a unit square by discretizing the square into a uniform grid with d^2 cells. The hypothesis space is the space of all possible uniform grids with d^2 cells for $d \in \{1, 2, 3, \dots\}$, where each cell gets a binary label.

We have a sample S of size n to learn the function. Let $\mathcal{H}_d$ be the hypothesis

set of uniform grids with d^2 cells. Let $\mathcal{H} = \bigcup_{d=1}^{\infty} \mathcal{H}_d$ be the hypothesis set of all possible uniform grids. Let $f(h)$ denote the number of cells in the hypothesis h. Let $d(h) = \sqrt{f(h)}$, then $d(h) \in \{1, 2, 3, \dots\}$, and $h \in \mathcal{H}_{d(h)}$.

1. Derive a generalization bound for learning with $\mathcal{H}_d$. (I.e., a bound on $L(h)$ that holds for all $h \in \mathcal{H}_d$ with probability at least $1 - \delta$.)
2. Derive a generalization bound for learning with $\mathcal{H}$. (I.e., a bound on $L(h)$ that holds for all $h \in \mathcal{H}$ with probability at least $1 - \delta$.)
3. Explain how to use the latter bound to select a prediction rule $h \in \mathcal{H}$.
4. What is the maximal number of cells as a function of n, for which your bound is non-vacuous? (It is sufficient to give an order of magnitude, you do not need to make the precise calculation.)
5. Explain how the density of the grid $d(h)$ affects the bound. Which terms in the bound (if any) increase as the density of the grid increases and which terms in the bound (if any) decrease as the density of the grid increases?

3 Early Stopping (30 points)

Early stopping is a widely used technique to avoid overfitting in models trained by iterative methods, such as gradient descent. In particular, it is used to avoid overfitting in training neural networks. In this question we analyze several ways of implementing early stopping. The technique sets aside a validation set S_{val}, which is used to monitor the improvement of the training process. Let $h_1, h_2, h_3, \dots$ be a sequence of models obtained after $1, 2, 3, \dots$ epochs of training a neural network or any other prediction model (you do not need to know any details about neural networks or their training procedure to answer the question). Let $\hat{L}(h_1, S_{\text{val}})$, $\hat{L}(h_2, S_{\text{val}})$, $\hat{L}(h_3, S_{\text{val}}), \dots$ be the corresponding sequence of validation errors on the validation set S_{val}.

1. Let h_{t^*} be the neural network returned after training with early stopping. In which of the following cases $\hat{L}(h_{t^*}, S_{\text{val}})$ is an unbiased estimate of $L(h_{t^*})$ and in which cases it is not? Please, explain your answer.
   (a) Predefined stopping: the training procedure always stops after 100 epochs and always returns the last model $h_{t^*} = h_{100}$.
   (b) Non-adaptive stopping: the training procedure is executed for a fixed number of epochs T, and returns the model h_{t^*} with the lowest validation error observed during the training process, i.e., $t^* = \arg \min_{t \in \{1,\dots,T\}} \hat{L}(h_t, S_{\text{val}})$.

(c) Adaptive stopping: the training procedure stops when no improvement in $\hat{L}(h_t, S_{\text{val}})$ is observed for a significant number of epochs. It then returns the best model observed ever during training. (This procedure is proposed in Goodfellow et al. (2016, Algorithm 7.1) or `https://www.quora.com/How-does-one-employ-early-stopping-in-TensorFlow`, but again, you do not need to know the details of the training procedure.)

2. Derive a high-probability bound (a bound that holds with probability at least $1 - \delta$) on $L(h_{t^*})$ in terms of $\hat{L}(h_{t^*}, S_{\text{val}})$, δ, and the size n of the validation set S_{val} for the three cases above. In the second case the bound may additionally depend on the total number of epochs T, while in the third case the bound may additionally depend on the index t^* of the epoch providing the optimal model. Please, solve the last case using the series $\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = 1$.[1]

3. The adaptive approach suggests stopping when "no improvement in $\hat{L}(h_t, S_{\text{val}})$ is observed for a significant number of epochs". A natural way of redefining the stopping criterion once we have the generalization bound is to stop when "no improvement in the generalization bound is observed for a significant number of epochs". The adaptive approach does not limit the number of epochs in advance, but what is the maximal number of epochs $T_{\max}$, after which it makes no sense to continue training according to the bound you derived in Point 2? Express $T_{\max}$ in terms of the number of validation samples n. It is sufficient to provide an order of magnitude of $T_{\max}$ in terms of n, you do not have to calculate the explicit constants.

4. How would your answer to the previous point change if you were to use the series $\sum_{i=1}^{\infty} \frac{1}{2^i} = 1$ for deriving the bound? (You should get that with these series you can run significantly less epochs in the adaptive approach compared to the series used in Point 2. Thus, unlike in the case of decision trees in the lecture notes, here the choice of the series has a significant impact.)

5. In this question we compare the adaptive procedure with non-adaptive. Assume that the two procedures use the same initialization, so that the corresponding models at epoch t are identical, and assume that the adaptive procedure has considered all the models h_t for $t \in \{1, \ldots, T_{\max}\}$. Let t^* be the index of the model h_{t^*} minimizing the adaptive bound and let T^* be the index of the model h_{T^*} minimizing the non-adaptive bound. Show that the generalization bound for adaptive stopping in Point 2 is never much worse

[1] We have $\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \sum_{i=1}^{\infty} \left(\frac{1}{i} - \frac{1}{i+1} \right) = 1$.

than the generalization bound for non-adaptive stopping, but in some cases the adaptive bound can be significantly lower.

Guidance: To simplify the analysis, throughout the question we assume that the confidence parameter $\delta \leq \frac{1}{2}$. For $T \geq 1$ it gives $\delta \leq \frac{1}{2} \leq \frac{T}{T+1}$.

(a) First, assume that $T \leq T_{\max}$. Let t^* be the index of the epoch selected by the adaptive procedure and T^* be the index of the epoch selected by the non-adaptive procedure. Since the adaptive procedure has selected t^* we know that the adaptive bound for epoch t^* is lower than the adaptive bound for epoch T^*. We also know that $T^* \leq T$, where T is the number of epochs in the non-adaptive approach. Use this information and do some bounding to show that for any confidence parameter $\delta \leq \frac{1}{2}$, the adaptive bound can be at most a multiplicative factor of $\sqrt{2}$ larger than the non-adaptive bound.

(b) [Optional] Now consider the case $T > T_{\max}$. Show that in this case the non-adaptive bound is at least $\frac{1}{\sqrt{2}}$. Since the losses are upper bounded by 1, any bound can be truncated at 1 and still be a valid bound. In other words, for any "bound" we can define a "truncated bound" = max(1, "bound") and it will still be a valid bound. So in this case the truncated adaptive bound also cannot exceed the non-adaptive bound by more than a multiplicative factor of $\sqrt{2}$.

(c) You have shown that under the assumption that $\delta \leq \frac{1}{2}$ the adaptive bound never exceeds the non-adaptive bound by more than a multiplicative factor of $\sqrt{2}$. Now provide *two* examples of sequences of empirical losses $\hat{L}(h_1, S_{\text{val}}), \hat{L}(h_2, S_{\text{val}}), \ldots$, for which the adaptive bound can be significantly smaller than the non-adaptive bound. In both cases you should have $T < T_{\max}$ and $\delta \leq \frac{1}{2}$.

(d) [Optional] Show that irrespective of the choice of T, there always exists a sequence of losses $\hat{L}(h_1, S_{\text{val}}), \hat{L}(h_2, S_{\text{val}}), \ldots$, for which $\frac{\text{adaptive bound}}{\text{non adaptive bound}} \leq \frac{1}{2}\left(\frac{\ln \frac{2}{\delta}}{n}\right)^{\frac{1}{4}}$.

Conclusion: depending on the data, the generalization bound for adaptive stopping can be significantly smaller than the generalization bound for non-adaptive stopping, and at the same time it is guaranteed that it is never worse by more than a multiplicative factor of $\sqrt{2}$.

4 Logistic Regression

done

4.1 Cross-entropy error measure (15 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).

done

4.2 Logistic regression loss gradient (15 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.

done

4.3 Log-odds (15 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.

References

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.