Machine Learning B

2021-2022

Home Assignment 3

Yevgeny Seldin **Marcus Holte Teller** **Sadegh Talebi**
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **14 December 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 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.
- 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. 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 The VC-dimension (50 points)

1. Let $\mathcal{H}$ be a finite hypothesis set with $|\mathcal{H}| = M$ hypotheses. Bound the VC-dimension of $\mathcal{H}$.

2. Let $\mathcal{H}$ be a hypothesis space with 2 hypotheses (i.e., $|\mathcal{H}| = 2$). Prove that $d_{\text{VC}}(\text{H}) = 1$.

3. What should be the relation between $d_{\text{VC}}(\text{H})$ and n in the VC generalization bound in Theorem 3.16 in Yevgeny's lecture notes in order for the bound to be non-trivial? [A bound on the loss that is greater than or equal to 1 is trivial, because we know that the loss is always bounded by 1. You do not have to make an exact calculation, giving an order of magnitude is sufficient.]

4. In the case of a finite hypothesis space, $|\mathcal{H}| = M$, compare the generalization bound that you can obtain with Theorem 3.16 in Yevgeny's lecture notes with the generalization bound in Theorem 3.2 in Yevgeny's lecture notes. In what situations which of the two bounds is tighter?

5. How many samples do you need in order to ensure that the empirical loss of a linear classifier selected out of a set of linear classifiers in $\mathbb{R}^{10}$ does not underestimate the expected loss by more than 0.01 with 99% confidence?

   Clarifications: (1) you are allowed to use the bound on the VC-dimension of linear classifiers mentioned in Yevgeny's lecture notes; (2) solving the question analytically is a bit tricky, you are allowed to provide a numerical solution. In either case (numerical or analytical solution), please, explain clearly in your report what you did.

6. Let $\mathcal{H}_+$ be the class of "positive" circles in $\mathbb{R}^2$ (each $h \in \mathcal{H}_+$ is defined by the center of the circle $c \in \mathbb{R}^2$ and its radius $r \in \mathbb{R}$; all points inside the circle are labeled positively and outside negatively). Prove that $d_{VC}(\mathcal{H}_+) \geq 3$.

7. Let $\mathcal{H} = \mathcal{H}_+ \cup \mathcal{H}_-$ be the class of "positive" and "negative" circles in $\mathbb{R}^2$ (the "negative" circles are negative inside and positive outside). Prove that $d_{VC}(\mathcal{H}) \geq 4$.

8. ***Optional question (0 points)*** Prove the matching upper bounds $d_{VC}(\mathcal{H}_+) \leq 3$ and $d_{VC}(\mathcal{H}) \leq 4$. [Doing this formally is not easy, but will earn you extra honor.]

9. What is the VC-dimension of the hypothesis space $\mathcal{H}_d$ of binary decision trees of depth d?

10. What is the VC-dimension of the hypothesis space $\mathcal{H}$ of binary decision trees of unlimited depth?

11. You have a sample of 100,000 points and you have managed to find a linear separator that achieves $\hat{L}_{\texttt{FAT}}(h, S) = 0.01$ with a margin of 0.1. Provide a bound on its expected loss that holds with probability of 99%. The input space is assumed to be within the unit ball and the hypothesis space is the space of linear separators.

12. **The fine details of the lower bound.** We have shown that if a hypothesis space $\mathcal{H}$ has an infinite VC-dimension, it is possible to construct a worst-case data distribution that will lead to overfitting, i.e., with probability at least $\frac{1}{8}$ it will be possible to find a hypothesis for which $L(h) \geq \hat{L}(h, S) + \frac{1}{8}$. But does it mean that hypothesis spaces with infinite VC-dimension are always deemed to overfit? Well, the answer is that it depends on the data distribution. If the data distribution is not the worst-case for $\mathcal{H}$, there may still be hope.

    Construct a data distribution $p(X, Y)$ and a hypothesis space $\mathcal{H}$ with infinite VC-dimension, such that for any sample S of more than 100 points with probability at least 0.95 we will have $L(h) \leq \hat{L}(h, S) + 0.01$ for all h in $\mathcal{H}$.

    Hint: this can be achieved with an extremely simple example.

2 The t-SNE Algorithm (50 points)

The aim of this exercise is to examine the t-SNE algorithm for visualization of some dataset and compare the result to that obtained by some classical method.

1. Implement the t-SNE algorithm.

2. Transform the data from `dataset1.txt` using both PCA and t-SNE. Report the choice of hyperparameters, step-size, if any. (You can use some available implementation for PCA.) Plot the data before transformation as well as the transformed data by both PCA and t-SNE.

3. Compare the plots of transformed data qualitatively. If there is a significant difference between the two, give a short explanation.