Machine Learning B

2024-2025

Home Assignment 4

Amartya Sanyal
Department of Computer Science
University of Copenhagen

The deadline for this assignment is **20 May 2025, 17:00**. 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 The Airline question

Question 1 (40 points).

1. An airline knows that any person making a 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.

   There are multiple ways to approach this question. We will guide you through two options. You are asked to solve the question in both ways.

   (a) Let p be the true probability of showing up for a flight (remember that p is unknown). In the first approach 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.

   (b) The second approach considers an alternative way of generating the two samples, using the same idea as in the proof of the VC-bound. Consider the following process of generating the two samples:

      i. We sample 10100 passenger show up events independently at random according to an unknown distribution p.

ii. And then we split them into 10000 passengers in the collected sample and 100 passengers booked for the 99-seats flight.

Bound the probability of observing a sample of 10000 with 95% show ups and a 99-seats flight with all 100 passengers showing up by following the above sampling protocol. If you do things right, you can get a bound of about 0.0062 (there may be some variations depending on how exactly you do the calculation).

2 PAC learnability

In this question, we will consider two variants of PAC learnability—positive/negative learnability and weak learnability—that appear to be weaker than the original PAC model considered in the lecture.

For a target concept class $\mathcal{C}$ and a target concept $c \in \mathcal{C}$, let $\mathcal{D}_c^+$ and $\mathcal{D}_c^-$ be arbitrary distributions over the instances labeled positively and negatively by c, respectively. Define the positive example oracle EX_c^+ as $\mathrm{Ex}\left(c; \mathcal{D}_c^+\right)$ and the negative example oracle EX_c^- as $\mathrm{Ex}\left(c; \mathcal{D}_c^-\right)$.

A concept class $\mathcal{C}$ is said to be *efficiently positively–negatively PAC learnable* by hypothesis class $\mathcal{H}$ if for every $\epsilon, \delta > 0$ there is a polynomial–time algorithm $\mathcal{A}$ which, given access to both EX_c^+ and EX_c^-, outputs a hypothesis $h \in \mathcal{H} \cup \{h_0, h_1\}$ satisfying with probability at least $1 - \delta$

$$\Pr_{x \sim \mathcal{D}_c^+}\left[\, h(x) = 0 \,\right] \le \epsilon \quad \text{and} \quad \Pr_{x \sim \mathcal{D}_c^-}\left[\, h(x) = 1 \,\right] \le \epsilon.$$

Question 2 (30 points)**.** Here h_0, h_1 are the always zero and the always one functions.

a) **Show** that if $\mathcal{C}$ is efficiently PAC learnable using $\mathcal{H}$ in the standard PAC model, then $\mathcal{C}$ is also efficiently positively–negatively PAC learnable using $\mathcal{H}$.

b) **Show** that if $\mathcal{C}$ is efficiently positively–negatively PAC learnable using $\mathcal{H}$, then $\mathcal{C}$ is also efficiently PAC learnable in the standard model.

3 Growth Function

Question 3 (30 points)**.** Solve the following three questions about growth function.

1. Let $\mathcal{H}$ be a finite hypothesis set with $|\mathcal{H}| = M$ hypotheses. Prove that $m_{\mathcal{H}}(n) \le \min\{M, 2^n\}$.

2. Let $\mathcal{H}$ be a hypothesis space with 2 hypotheses (i.e., $|\mathcal{H}| = 2$). Prove that $m_{\mathcal{H}}(n) = 2$. (Put attention that you are asked to prove the equality, $m_{\mathcal{H}}(n) = 2$, not an inequality.)

3. Prove that $m_{\mathcal{H}}(2n) \le m_{\mathcal{H}}(n)^2$.

4 PAC Learning 3-CNFs [Optional]

Question 4 (0 points)**.** A *3-CNF formula* over Boolean variables $x_1, \dots, x_d$ is a conjunction of clauses, each clause being a disjunction of at most three literals (where a literal is either x_i or $\neg x_i$). Let

$$\mathsf{CNF}_3 = \Big\{ \bigwedge_{j=1}^{k} C_j \;\Big|\; C_j = (\ell_{j,1} \vee \ell_{j,2} \vee \ell_{j,3}),\ \ell_{j,*} \in \{x_i, \neg x_i\} \Big\}.$$

Prove that CNF_3 is PAC learnable in time polynomial in d, $1/\epsilon$, and $\ln(1/\delta)$.

Argue that with the usual sample size for VC dimension $\Theta(d^3)$, one can enumerate all candidate 3-clauses consistent with the sample and produce a 3-CNF hypothesis efficiently.

5 Computational Hardness of learning [Optional]

In this exercise we will look at some more hardness results in learning theory including for the case of agnostic learning. We will try to follow a similar proof structure as in the lecture.

We will require the knowledge of the following *hard* problem.

Definition 1 (Zero–One Integer Programming (ZIP))**.** This problem is known to be NP-COMPLETE. An instance of ZIP consists of a matrix $A \in \{0,1\}^{s \times d}$, a vector $b \in \{0,1\}^s$, and a pair (c, B) with $c \in \{0,1\}^d$ and $B \in \mathbb{Z}_+$. The decision problem is to determine whether there exists a Boolean assignment to the d variables $z_1, \dots, z_d$ such that

$$\Big(\forall\, 1 \le i \le s : \sum_{j=1}^{d} A_{i,j}\, z_j \;\le\; b_i\Big) \quad \text{and} \quad \sum_{j=1}^{d} c_j\, z_j \;\ge\; B.$$

Question 5 (0 points)**.** Now consider the following problem of learning Boolean threshold functions

Let $\mathcal{X}_d = \{0,1\}^d$ and for each $w \in \{0,1\}^d$ and $k \in \mathbb{N}$ define

$$f_{w,k} \colon \mathcal{X}_d \to \{0,1\}, \qquad f_{w,k}(x) = \begin{cases} 1, & \text{if } \sum_{i=1}^d w_i x_i \ \geq k, \\ 0, & \text{otherwise.} \end{cases}$$

Let

$$\mathrm{TH}_d = \left\{ f_{w,k} \mid w \in \{0,1\}^d,\ 0 \leq k \leq d \right\}, \quad \mathrm{TH} = \bigcup_{d \geq 1} \mathrm{TH}_d.$$

Show that, unless RP = NP, there is no efficient PAC-learning algorithm for learning TH if the output hypothesis is required to lie in TH (i.e. a proper learner).

Hint: Use the ZIP problem to show hardness. Follow the same proof strategy as for DNF-hardness: construct a dataset with exactly 2 positive examples, $s+B+2$ negative examples, each in dimension $d+B+2$, and show consistency of a threshold in TH corresponds to a solution of the ZIP instance.