Cairo University - Faculty of Engineering
Computer Engineering Department
CMP4040 – Machine Learning
Spring 2023

Assignment #1

Solve the following exercises and report your answers.

Exercise 1.10 (Code + Report)

Here is an experiment that illustrates the difference between a single bin and multiple bins. Run a computer simulation for flipping 1,000 fair coins. Flip each coin independently 10 times. Let's focus on 3 coins as follows: c_1 is that first coin flipped; c_{rand} is a coin you choose at random; c_{min} is the coin that has the minimum frequency of heads (pick the earlier one in case of a tie). Let v_1, v_{rand} and v_{min} be the fraction of heads you obtain for the respective three coins.

(a) What is mu for the three coins selected?
(b) Repeat the entire experiment a large number of times (e.g., 100,000 runs of the entire experiment) to get several instances of v_1, v_{rand} and v_{min} and plot the histograms of the distributions of v_1, v_{rand} and v_{min}. Notice that which coins end up being c_{rand} and c_{min} may differ from one run to another.
(c) Using (b), plot estimates of $P[|\nu - \mu| > \epsilon]$ as a function of ϵ, together with the Hoeffding bound $2e^{-2\epsilon^2 N}$ (on the same graph).
(d) Which coins obey the Hoeffding bound, which ones do not? Explain why.
(e) Relate part (d) to the multiple bins in Figure 1.

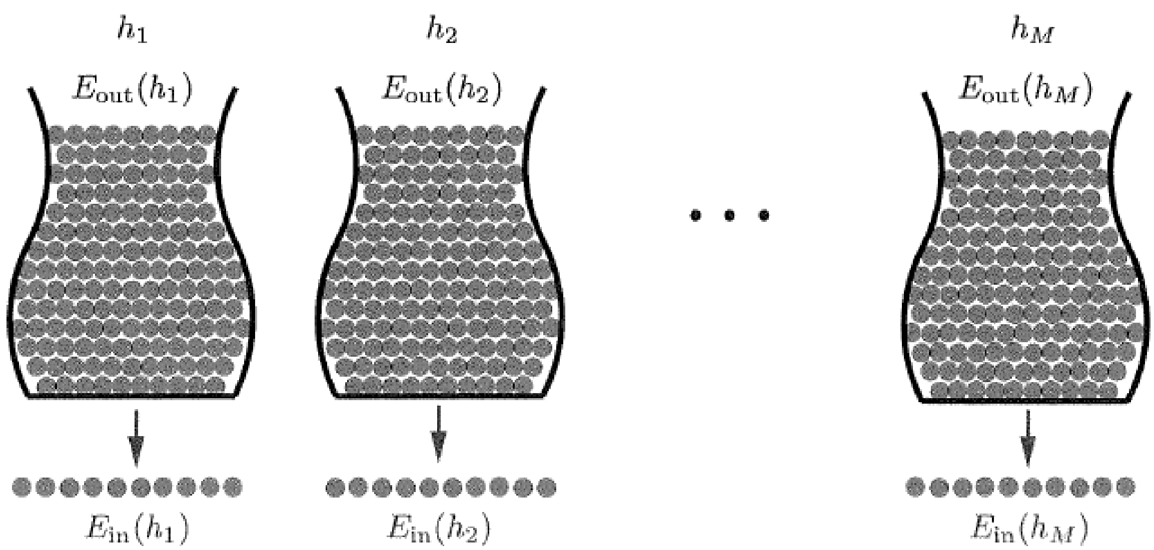

Figure 1 Multiple bins depict the learning problem with M hypotheses

Problem 1.4 (Code + Report)

This problem leads you to explore the perceptron learning algorithm with data sets of different sizes and dimensions. You will have to generate random dataset as described in Exercise 1.4 of the book which works as follows: For a specific number of dimensions *d*, Choose a random target function *f* (randomly pick values for the *d+1* weights of the perceptron), then generate *N* random data points $\left\{x_i \in R^d \mid i\in[1, N]\right\}$ and label each data point $y_i = f(x_i)$ according to the selected target function *f*.

(a) Generate a linearly separable data set of size 20 where *d=2*. Plot the examples $\{x_n\}$ as well as the target function *f* on a plane. Be sure to mark the examples from different classes *y* differently and add labels to the axes of the plot.
(b) Run the perceptron learning algorithm on the data set above. Report the number of updates that the algorithm takes before converging. Plot the examples, the target function *f* and the final hypothesis *g* in the same figure. Comment on whether *f* is close to *g*.
(c) Repeat everything in (b) with another randomly generated data set of size 20. Compare your results with (b).
(d) Repeat everything in (b) with another randomly generated data set of size 100. Compare your results with (b).
(e) Repeat everything in (b) with another randomly generated data set of size 1,000. Compare your results with (b).
(f) Modify the algorithm such that it takes d=10 instead of 2. Randomly generate a linearly separable data set of size 1,000 and feed the data set to the algorithm. How many updates does the algorithm take to converge?
(g) Repeat the algorithm on the same data set as (f) for 100 experiments. In the iterations of each experiment, pick *x(t)* randomly instead of deterministically. Plot a histogram for the number of updates that the algorithm takes to converge.
(h) Summarize your conclusions with respect to accuracy and running time as a function of *N* and *d*.

Problem 1.6 (Handwritten)

Consider a sample of 10 marbles drawn independently from a bin that holds red and green marbles. The probability of a red marble is μ. For μ = 0.05, μ = 0.5, and μ = 0.8, compute the probability of getting no red marbles (ν = 0) in the following cases.

(a) We draw only one such sample. Compute the probability that ν = 0.
(b) We draw 1,000 independent samples. Compute the probability that (at least) one of the samples has ν = 0.
(c) Repeat (b) for 1,000,000 independent samples.

Problem 2.5 (Handwritten)

Prove by induction that $\sum_{i=0}^{D} (N\ i) \leq N^{D} + 1$, hence $m_{H}(N) \leq N^{dvc} + 1$

Rules:

- This is an individual assignment.
- Any evidence of plagiarism will be penalized by all the grades of the assignment.
- Beside the name of each problem, you will find the expected solution format which can be either of the following:
    - **(Handwritten)**: These problems should be solved on paper and scanned. The handwriting and the scanning result should both be clear.
    - **(Code + Report)**: These problems require coding to solve. In that case, you should write the needed code, generate the results, and solve the problem accordingly. You will find an attached python notebook. It's required to update the notebook with your solutions and comments (in the allocated areas). You should **only** submit the updated notebook (with solutions displayed clearly) with your report.

Deliverables: One report of your solutions (PDF) + One python notebook with your code and solutions displayed clearly. Remember to include your name, section & bench number in the first page of the report.

Deadline: Wednesday, March 15th, 2023, 23:59.