# CSC 588 Spring 2024: Homework 2

Kwang-Sung Jun

Due: Feb 20, 2024, 11:59pm MST

Please complete the following exercises and read the following instructions carefully.

- Your solutions to these problems will be graded based on both correctness and clarity. Your arguments should be clear: there should be no room for interpretation about what you are writing. Otherwise, I will assume that they are wrong, and grade accordingly.
- If you feel unable to make progress on any of the questions, you can post your questions on Piazza. Try posing your questions to be as general as possible, so that it can promote discussion among the class.
- You are encouraged to discuss the homework questions with your classmates, but the discussions should only be at a high level, and you should write your solutions in your own words. For every question you have had discussions on, please mention explicitly whom you have discussed with; otherwise it may be counted as academic integrity violation.
- For detailed homework policies, please read the course syllabus carefully, available on the course website.
- **Show all work along with answers to get the full credit**.
- Place your final answer into an 'answer box' that can be easily identified (unless the answer is a proof).
- There will be no late days. Late homework result in zero credit. Not even one minute. It is a good idea to set yourself up your own deadline like one day before it is due.
- Each subproblem (i.e., Problem X.Y) is worth 10 points.
- This calibration homework counts toward total homework grades (10 pts / 40 pts).

Submission instruction:

- Submit homework via gradescope. You can hand-write your answers and scan them to make it a PDF, or type up your answers as pdf using LaTeX. If you use your phone camera, I recommend using TurboScan (smartphone app) or similar ones to avoid uploading a slanted image or showing the background. Make sure you rotate it correctly.
- Watch the video and follow the instruction for the submission: `https://youtu.be/KMPoby5g_nE`
- Report the code as part of the answer as texts. You should also submit the code to a separate submission entry in gradescope 'HW# - code' as well.

# Problem 1

Given a random variable $X$ with expectation $\mu$, prove that for any $w \in \mathbb{R}$,

$$\mathbb{E}\left[(X-w)^2\right] = \mathbb{E}\left[(X-\mu)^2\right] + (\mu - w)^2.$$

(Note that this implies that $\mu = \operatorname{argmin}_{w \in \mathbb{R}} \mathbb{E}\left[(X-w)^2\right]$.)

# Problem 2

Consider the cost-sensitive classification (CSC) problem: each example is represented by a pair $(z, c)$, where $z \in \mathcal{Z}$ is its feature part, and $c = (c(1), ..., c(K)) \in [0, M]^K$ is its cost vector, where for $k \in [K] := \{1, \dots, K\}$, $c(k)$ represents the cost of predicting the example with class $k$. Given a classifier $h : \mathcal{Z} \to [K]$, the cost of $h$ on example $(z, c)$ is defined as $c(h(z))$. The performance of a classifier $h$ on a distribution $D$ is measured by its expected cost $L(h, D) = \mathbb{E}_{(z,c)\sim D}\left[c(h(z))\right]$.

(a) Given a binary classification example $(x, y)$ (where $y \in \{1, 2\}$), how can we construct a CSC example $(z, c)$, such that $\mathbb{1}(h(x) \neq y)$, the 0-1 classification error indicator of a binary classifier $h$ on $(x, y)$, equals its cost on $(z, c)$?

(b) Suppose we have a finite hypothesis class $\mathcal{H}$, and are given a set $S$ of $m$ CSC examples $(z_1, c_1), \dots, (z_m, c_m)$ drawn iid from $D$. Define the ERM as $\hat{h} = \operatorname{argmin}_{h \in \mathcal{H}} L(h, S)$, where $L(h, S) = \frac{1}{m}\sum_{i=1}^m c_i(h(z_i))$. For any $\delta > 0$, give an upper bound on $L(\hat{h}, D) - \min_{h' \in \mathcal{H}} L(h', D)$, that holds with probability $1 - \delta$ and goes to zero with $m \to \infty$. Justify your answer.

# Problem 3 (20 points)

Suppose we have an algorithm $\mathcal{B}$ that learns hypothesis class $\mathcal{H}$ in the following sense. There exists a function $g : (0, 1) \to \mathbb{N}$, such that for any distribution $D$, for any $\epsilon > 0$, if $\mathcal{B}$ draws $m \geq g(\epsilon)$ iid training examples from $D$, then with probability at least $\frac{1}{10}$, $\mathcal{B}$ returns a classifier $\hat{h}$ whose excess (generalization) error on $D$ $(\operatorname{err}(\hat{h}, D) - \min_{h \in \mathcal{H}} \operatorname{err}(h, D))$ is at most $\epsilon$.

Now, given $\mathcal{B}$, and the ability to draw fresh training examples from the data distribution, how can you design an algorithm $\mathcal{A}$ that $(\epsilon, \delta)$-agnostic PAC learns $\mathcal{H}$ for any $\epsilon, \delta > 0$? What is your $\mathcal{A}$'s sample complexity?
Hint 1: You may want to consider the algorithm below (you will have to find the right value of $k$ and $m$ to answer your question).
Hint 2: You may want to consider the event $E_1 = \{\exists i \in [k] : \operatorname{err}(h_i, D) \leq \min_{h \in \mathcal{H}} \operatorname{err}(h, D) + \frac{\epsilon}{2}\}$.

**Algorithm 1** An agnostic PAC algorithm $\mathcal{A}$ for $\mathcal{H}$

```
Require: Target error $\epsilon$, failure probability $\delta$, parameters $k$ and $m$.
1: Repeatedly run $\mathcal{B}(\frac{\epsilon}{2})$ for $k$ times, getting classifiers $h_1, \dots, h_k$.
2: $S \leftarrow$ Sample $m$ iid examples from $D$.
3: Select $\hat{h} = \arg\min_{h \in \hat{\mathcal{H}}} \widehat{\operatorname{err}}(h, S)$, where $\hat{\mathcal{H}} = \{h_1, \dots, h_k\}$.
```

# Problem 4

Let us consider the threshold class:

$$\mathcal{H} = \{h_\theta(x) = 2 \cdot \mathbb{1}\{x \geq \theta\} - 1 : \theta \in (0, 1)\}\ .$$

Assume that there exists $h^* \in \mathcal{H}$ such that $y = h^*(x)$ where $(x, y) \sim \mathcal{D}$; i.e., realizable setting. As usual, we work with the i.i.d. data $\{(x_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m$. Assume $x_i \sim \text{Uniform}[0, 1]$. The purpose of this problem is to empirical verify that the generalization bound for the ERM $\widehat{h}_m$ is

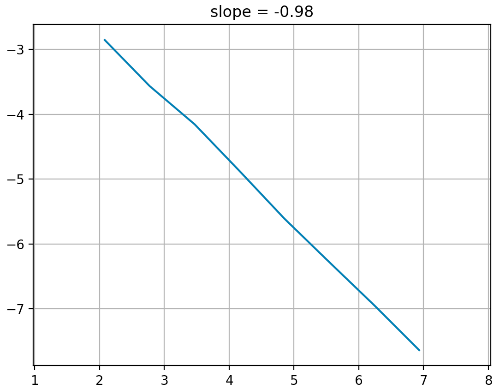
$$\text{err}(\widehat{h}_m, \mathcal{D}) \approx c/m$$

for some $c > 0$. This constant $c$ may depend on some important quantities like VC dimension of $\mathcal{H}$, but we are ignoring them for now and focusing on how it scales with $m$. The above implies that $\ln(\text{err}(\widehat{h}_m, \mathcal{D})) \propto -\ln(m)$. Therefore, if we plot $\ln(\text{err}(\widehat{h}_m, \mathcal{D}))$ as a function of $\ln(m)$, then we will see a straight line with slope -1. Our goal is to show verify this by simulation.

**(a)** Implement the following pseudocode:
Repeat the following 200 times:

- For each $m \in \{2^3, 2^4, \ldots, 2^{10}\}$
  - Sample $\{(x_i, y_i)\}_{i=1}^m \sim \mathcal{D}^m$.
  - Compute the ERM $\widehat{h}_m$ (break ties with any rule of your choice).
  - Compute $\text{err}(\widehat{h}_m, \mathcal{D})$ (do this in a closed form expression).

**(b)** Plot the average and the slope.
Assume that the label is generated by $h_\theta$ with $\theta = \frac{1}{2}$. For every $m$, compute the average $\text{err}(\widehat{h}_m, \mathcal{D})$ over 200 trials and call it $z_m$. Plot $\ln(z_m)$ as a function of $\ln(m)$. To compute the slope, perform linear regression on $\{(\ln(m), \ln(z_m))\}$ and report the slope in the title. As an example, here is what I obtained.



# Problem 5

How long did it take to complete your homework?