CSC 588 Spring 2022: Homework 1 (Calibration)

Chicheng Zhang

January 2022

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.

This homework is due on ~~Jan 20, 2022~~ Jan 24, 2022, 5pm MST, on gradescope. Note that this calibration homework counts toward total homework grades (4 pts / 40 pts).

Problem 1

Denote by $\mathrm{B}(n, p)$ the binomial distribution with n being the number of trials, and p being the success probability of each trial. Suppose Y is a random variable such that $\mathbb{P}(Y = +1) = \mathbb{P}(Y = -1) = \frac{1}{2}$. In addition, X has the following conditional probability distribution given Y: given $Y = -1$, $X \sim \mathrm{B}(3, \frac{2}{3})$; given $Y = +1$, $X \sim \mathrm{B}(2, \frac{1}{3})$. Answer the following questions:

1. Calculate the joint probability table of (X, Y).
2. What is the value of $\mathbb{P}(Y = -1 \mid X = 1)$?
3. Suppose we would like to find a function $f : \{0, 1, 2, 3\} \to \{-1, +1\}$ that minimizes its *classification error* $\mathbb{P}(f(X) \neq Y)$. Can you find the optimal f, and what is the optimal value of classification error?

Problem 2

Suppose we have a deterministic set of examples $x_1, \ldots, x_n \in \mathbb{R}^d$, a deterministic vector $\theta \in \mathbb{R}^d$, and a set of independent random variables (noise) $\epsilon_1, \ldots, \epsilon_n$, where for each i, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ (here N denotes the normal distribution). Each example x_i is associated with a *label* y_i, defined by $y_i = \langle \theta, x_i \rangle + \epsilon_i$. Assume that $\Sigma = \sum_{i=1}^n x_i x_i^\top$ is invertible. Answer the following questions:

1. What is the joint distribution of $(y_1, \ldots, y_n)$?

2. Define random vector $\hat{\theta} = \Sigma^{-1}(\sum_{i=1}^{n} x_i y_i)$ (this is the *ordinary least squares* estimator). What is the distribution of $\hat{\theta}$?

3. Given a deterministic vector v, what is the distribution of random variable $\left\langle v, \hat{\theta} - \theta \right\rangle$? Find a decreasing function $f : (0, 1] \to \mathbb{R}$, so that the statement

$$\forall \delta \in (0, 1] \,.\, \mathbb{P}\left(\left| \left\langle v, \hat{\theta} - \theta \right\rangle \right| \geq f(\delta) \right) \leq \delta$$

holds. (You are free to use e.g. Markov's Inequality, Chebyshev's Inequality, or other probability inequalities you know to construct your f; the tightness of function f won't be graded.)

Problem 3

In this exercise, we verify the theoretical claims in the Perceptron convergence theorm empirically. Throughout, we assume that the sequence of online classification examples $(x_1, y_1), \ldots, (x_n, y_n)$ is such that (1) for all t, $\|x_t\| \leq 1$; (2) there exists some $w^\star$, such that $\|w^\star\| \leq 1$, and for all t, $y_t \langle w^\star, x_t \rangle \geq \gamma$.

In the class, we have seen that the Perceptron algorithm always makes at most $1/\gamma^2$ mistakes throughout the process. In this exercise, we verify this claim empirically.

1. Let $w^\star = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$. Write a function `generate_data` that receives a sample size parameter n and margin parameter γ as input, and output n independently drawn examples $(x_1, y_1), \ldots, (x_n, y_n)$, such that for each i, x_i comes from the uniform distribution over the region $R_\gamma = \left\{ x \in \mathbb{R}^2 : \|x\|_2 \leq 1, |\langle w^\star, x \rangle| \geq \gamma \right\}$, and $y_i = \text{sign}(\langle w^\star, x_i \rangle)$.

   Run `generate_data`($\texttt{n} = 1000, \gamma = 1/32$), give a scatterplot of the output examples in a 2-dimensional plane, where for every example, its location indicates its x value, and its color indicates its y value.

2. In the Perceptron algorithm, define $w^{(m)}$ to be the linear classifier weight iterate after m mistakes are made; for example, $w^{(0)} = (0, 0)$. Also let M be the total number of mistakes Perceptron makes throughout the processing of all n examples.

   Write a function `sim_perceptron` that takes a linearly separable dataset of size n as input, simulates the execution of Perceptron by processing the n examples from this dataset one by one, and outputs M. The function should also generate two plots: $\left\langle w^{(m)}, w^\star \right\rangle$, $\|w^{(m)}\|$ as functions of m, for $m \in \{0, 1, \ldots, M\}$.

   Now run `sim_perceptron` on the dataset you generate from step 1; report the M value, the final weight w_{n+1}, and show the two plots. Use the plots to verify that

$$\forall m \in \{0, 1, \ldots, M\} \,.\, \left\langle w^{(m)}, w^\star \right\rangle \geq \gamma m, \text{ and } \|w^{(m)}\| \leq \sqrt{m}.$$

3. For each value of $\gamma \in \left\{ 2^{-i} : i \in \{1, \ldots, 6\} \right\}$, do the following:

   (a) Repeatedly run `generate_data`($\texttt{n} = 100, \gamma$) 10 times to generate 10 datasets.

   (b) Run `sim_perceptron` on the 10 datasets, obtaining 10 output values $M_{\gamma,1}, \ldots, M_{\gamma,10}$.

   (c) Compute the average value $\bar{M}_\gamma = \frac{1}{10} \sum_{j=1}^{10} M_{\gamma,j}$.

   Now, plot $\bar{M}_\gamma$ as a function of γ. Is your plot of $\bar{M}_\gamma$ always below the plot of the function $g(\gamma) = \frac{1}{\gamma^2}$? Why?