CSC 588 Spring 2021: Calibration Homework

Chicheng Zhang

January 2021

Please complete the following set of exercises on your own. This homework is due on Jan 15, 2021.

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 $y_i = \langle \theta, x_i \rangle + \epsilon_i$. Denote by $\Sigma = \sum_{i=1}^n x_i x_i^\top$. Answer the following questions:

1. What is the joint distribution of $(y_1, \ldots, y_n)$?
2. Define random vector $\hat{\theta} = \Sigma^{-1}(\sum_{i=1}^n x_i y_i)$. 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 function $f : \mathbb{R} \to \mathbb{R}$, so that the statement that

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

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

Problem 3

In the class, we have seen that the Perceptron algorithm, when receiving a sequence of examples that are linearly separable by a margin γ [1] as input, makes at most $1/\gamma^2$ mistakes throughout the process. In this exercise, we verify this claim empirically. Throughout, we assume $w^\star = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$.

[1] More precisely, we require the sequence of examples $(x_1, y_1), \ldots, (x_n, y_n)$ to be 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, $y_t \langle w^\star, x_t \rangle \geq \gamma$.

1. 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 $\texttt{generate\_data}(\texttt{n} = \texttt{100}, \gamma = \texttt{0.25})$, 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. Given a sequence of examples $(x_1, y_1), \ldots, (x_n, y_n)$ linearly separable by a margin γ, consider running the Perceptron algorithm forever by cycling through the dataset (more precisely, at time step 1, (x_1, y_1) is shown; subsequently, if at time step t, example (x_i, y_i) is shown, then at time step $t + 1$, $(x_{(i \bmod n)+1}, y_{(i \bmod n)+1})$ will be shown). Building on the Perceptron algorithm, show that it is possible to write a program `cycling_perceptron_mistakes` that calculates the total number of mistakes Perceptron makes on this infinite cycling sequence. (Describing your implementation in words would suffice; presenting your code is welcome but not required).

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

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

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

   (c) Compute the average value $\hat{m}_\gamma = \frac{1}{10} \sum_{j=1}^{10} m_{\gamma,j}$.

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