Boston University
Department of Electrical and Computer Engineering

ENG EC 414 Introduction to Machine Learning

HW 2 Solution

Issued: Tue 8 Sept 2020 **Due:** 4:55pm Tue 15 Sept 2020 on Gradescope

Important: Before you proceed, please read the documents pertaining to *Homework formatting and submission guidelines* in the Homeworks section of Blackboard.

Note: Problem difficulty = number of coffee cups ☕

Problem 2.1 [8pts] *(max, argmax, min, argmin)* Let $f(x) := (2 + \sin(2\pi x))$ and $\mathcal{A} := [0, 2]$.

(a) [2pts] Compute: $\max_{x\in\mathcal{A}} f(x)$ and $\operatorname*{argmax}_{x\in\mathcal{A}} f(x)$.

(b) [2pts] Compute: $\min_{x\in\mathcal{A}} f(x)$ and $\operatorname*{argmin}_{x\in\mathcal{A}} f(x)$.

(c) [2pts] Compute: $\operatorname*{argmax}_{x\in\mathcal{A}} \left[2 \cdot e^{-f(x)+5} - 7\right]$.

(d) [2pts] Compute: $\operatorname*{argmin}_{x\in\mathcal{A}} \left[11 - \ln(2 \cdot f(x) + 5)\right]$.

Solution: Since $\sin(2\pi z)$ is periodic with period 1 so is the function $f(z) := (2 + \sin(2\pi z))$. The interval $\mathcal{A} := [0, 2]$ has a length which is two times the period of $f(z)$. Thus, the values of the function $f(z)$ repeat two times over $\mathcal{A}$. Moreover, the sine function attains the maximum value of $+1$ at 0.5π and at points separated from 0.5π by an integer multiple of 2π. Similarly, the sine function attains a minimum value of -1 at -0.5π and at points separated from -0.5π by an integer multiple of 2π. With these preliminary observations we are ready to compute the quantities asked in the problem.

(a) [2pts] $\boxed{\text{Maximum value} = 2 + 1 = 3 \text{ at } z = 0.25 \text{ and } z = 1 + 0.25 = 1.25.}$.

(b) [2pts] $\boxed{\text{Minimum value} = 2 - 1 = 1 \text{ at } z = -0.25 + 1 = 0.75 \text{ and } z = 1 + 0.75 = 1.75.}$.

(c) [2pts] $\boxed{\operatorname*{argmin}_{z\in\mathcal{A}} f(z) = \{0.75, 1.75\}}$. The function $\phi(u) := \left[2 \cdot e^{-u+5} - 7\right]$ is a strictly decreasing function because its first derivative is $\frac{d}{du}\phi(u) = -2 \cdot e^{-u+5}$ which is strictly negative for all $u \in \mathbb{A}$. Thus $\operatorname*{argmax}_{z\in\mathcal{A}} \phi(f(z)) = \operatorname*{argmin}_{z\in\mathcal{A}} f(z)$.

(d) [2pts] $\boxed{\operatorname*{argmax}_{z\in\mathcal{A}} f(z) = \{0.25, 1.25\}}$. The function $\psi(u) := [11 - \ln(2 \cdot u + 5)]$ is a strictly decreasing function because its first derivative is $\frac{d}{du}\psi(u) = -\frac{2}{2u+5}$ which is strictly negative for all $u \in \mathbb{A}$. Thus $\operatorname*{argmin}_{z\in\mathcal{A}} \psi(f(z)) = \operatorname*{argmax}_{z\in\mathcal{A}} f(z)$.

Problem 2.2 [16pts] *(Training errors for constant predictors)* A dataset has $n = 12$ examples with feature vectors in $\mathbb{R}^{10}$ and the following labels $\{y_1 = 2.0, y_2 = 1.0, y_3 = 5.0, y_4 = 3.0, y_5 = 2.0, y_6 = 4.0, y_7 = 1.0, y_8 = 4.0, y_9 = 6.0, y_{10} = 4.0, y_{11} = 2.0, y_{12} = 6.0\}$. We consider constant predictors with different loss functions.

(a) [4pts] *(Empirical average zero-one loss)*
Let $F_{\text{error}}(y) := \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[y \neq y_j]$, where $\mathbf{1}[A]$ is 1 if A is true and 0 otherwise. Compute the argmin.

(b) [5pts] *(Empirical average squared-error loss)*
Let $F_{\text{MSE}}(y) := \frac{1}{n}\sum_{j=1}^{n}(y - y_j)^2$. Compute the argmin.

(c) [7pts] ☕☕ *(Empirical average absolute-error loss)*
Let $F_{\text{MAE}}(y) := \frac{1}{n}\sum_{j=1}^{n} |y - y_j|$. Compute the argmin.

Solution:

(a) [4pts] $\boxed{\{2.0, 4.0\}}$

$$\begin{aligned}
\operatorname*{argmin}_{y\in\mathbb{R}} F_{\text{error}}(y) &= \operatorname*{argmin}_{y\in\mathbb{R}} \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[y \neq y_j] \\
&\overset{(i)}{=} \operatorname*{argmin}_{y\in\mathbb{R}} \frac{1}{n}\sum_{j=1}^{n} \Big(1 - \mathbf{1}[y = y_j]\Big) \\
&= \operatorname*{argmin}_{y\in\mathbb{R}} \frac{1}{n}\left(n - \sum_{j=1}^{n} \mathbf{1}[y = y_j]\right) \\
&\overset{(ii)}{=} \operatorname*{argmin}_{y\in\mathbb{R}} \left(1 - \frac{n_y(\mathbf{y})}{n}\right) \\
&\overset{(iv)}{=} \{2.0, 4.0\}\,.
\end{aligned}$$

Explanation of numbered steps: *(i)* The event $\{y \neq y_j\}$ occurs if, and only if (iff), the complement event $\{y = y_j\}$ does not occur. *(ii)* There are 12 examples, but they only take 6 different values: $\{1.0, 2.0, 3.0, 4.0, 5.0, 6.0\}$. If we let $n_y(\mathbf{x})$ denote the number of examples whose value equals y, then $n_1(\mathbf{x}) = 2, n_2(\mathbf{x}) = 3, n_3(\mathbf{x}) = 1, n_4(\mathbf{x}) = 3, n_5(\mathbf{x}) = 1, n_6(\mathbf{x}) = 2$. Note that the sum of all these counts must equal n, the total number of examples, since every example is accounted for in one of these counts. *(iii)* Property of argmax and argmin: the values of y which minimize $1 - (n_y(\mathbf{x})/n)$ are precisely the values of y which maximize $n_y(\mathbf{x})/n$. *(iv)* The labels $y = 2.0$ and $y = 4.0$ both have the same maximum count of 3, i.e., $n_2(\mathbf{x}) = n_4(\mathbf{x}) = 3$.

(b) [5pts] $\boxed{\operatorname*{argmin}_{y\in\mathbb{R}} F_{\text{MSE}}(y) = \frac{1}{n}\sum_{j=1}^{n} y_j = \sum_{l=1}^{6} l \times \frac{n_l(\mathbf{x})}{n} = \frac{1\times2+2\times3+3\times1+4\times3+5\times1+6\times2}{12} = \frac{40}{12} = 3.33}$

(c) [7pts] $\boxed{\operatorname*{argmin}_{y\in\mathbb{R}} F_{\text{MAE}}(y) = [3.0, 4.0]}$

Note: $[3.0, 4.0]$ is a continuous range of real numbers from 3.0 to 4.0.

The function $g(t) = |t - a|$ as a function of t is continuous over the entire real line and is differentiable everywhere except at $t = a$ where its derivative is undefined. Specifically, we have

$$\frac{d}{dt}g(t) = \begin{cases} +1 & t > a \\ -1 & t < a \\ \text{undefined} & t = a. \end{cases}$$

For $t \neq a$ this can be more compactly written as $\frac{d}{dt}g(t) = \operatorname{sign}(t-a)$.

Thus $F_{\mathrm{MAE}}(y|\mathbf{x}) = {}^1\!/\!{}_n \sum_{j=1}^{n} |y - y_j|$ as a function of y is also continuous everywhere on $\mathbb{R}$ and differentiable everywhere except at $y = y_j, j = 1, \ldots, n$. For all $y \notin \{y_1, \ldots, y_n\}$, the first derivative of $F_{\mathrm{MAE}}(y|\mathbf{x})$ with respect to the variable y is given by

$$\begin{aligned} \frac{d}{dy} F_{\mathrm{MAE}}(y|\mathbf{x}) &= \frac{1}{n} \sum_{j=1}^{n} \frac{d}{dy} |y - y_j| \\ &= \frac{1}{n} \sum_{j=1}^{n} \operatorname{sign}(y - y_j) \\ &= \frac{1}{n} \left(\left| \{ j \in \{1, \ldots, n\} : y_j < y \} \right| - \left| \{ j \in \{1, \ldots, n\} : y_j > y \} \right| \right), \end{aligned}$$

i.e., the fraction of examples with value less than y minus the fraction of examples with value greater than y. Note that |set| denotes the number of elements in the set. Observe that for any value $y \in (3, 4)$ the number of labels $< y$ equals the number of labels $> y$. For any value $y < 3$, the number of labels $< y$ is strictly smaller than the number of labels $> y$. For any value $y > 4$, the number of labels $< y$ is strictly larger than the number of labels $> y$. Thus $\widehat{L}_{\mathrm{MAE}}(y|\mathbf{x})$ is strictly decreasing for all $y < 3$, strictly increasing for all $y > 4$, and constant for all $3 < y < 4$. Since the function is also continuous, it attains its minimum value at $[3, 4]$ (not just $(3, 4)$).

In general, the MAE is minimized at any **median** value $\widehat{\eta}_{\text{label}}$ of all the labels. Loosely speaking, the median is a "middle" value: the number of values to the left and right of the median are equal. More generally since there may not be an exact middle value, the median is defined as any number $\widehat{\eta}_{\text{label}}$ which satisfies the following two conditions: (i) $\left|\{j : y_j < \widehat{\eta}_{\text{label}}\}\right| \leq \left|\{j : y_j \geq \widehat{\eta}_{\text{label}}\}\right|$ and (ii) $\left|\{j : y_j \leq \widehat{\eta}_{\text{label}}\}\right| \geq \left|\{j : y_j > \widehat{\eta}_{\text{label}}\}\right|$. In our particular problem any real number in the range $[3.0, 4.0]$ is a median value according to this definition.

Problem 2.3 [12pts] *(Argmin for random variables)*

(a) [5pts] ☕ *(Expected squared-error loss given* **x***)* Suppose $p(y|\mathbf{x})$ denotes the conditional pdf/pmf of random variable Y (modeling labels) given that random vector $\mathbf{X}$ (modeling feature vectors) equals $\mathbf{x}$. Let $L_{\mathrm{MSE}}(y|\mathbf{x}) := E[(Y - y)^2 \mid \mathbf{X} = \mathbf{x}]$, i.e., the conditional expectation of the squared-error loss given $\mathbf{X} = \mathbf{x}$. Express $h(\mathbf{x}) := \operatorname*{argmin}_{y \in \mathbb{R}} L_{\mathrm{MSE}}(y|\mathbf{x})$ in terms of $p(y|\mathbf{x})$.

(b) [7pts] ☕☕ *(Expected absolute-error loss given* **x***)* Suppose $p(y|\mathbf{x})$ denotes the conditional pdf/pmf of random variable Y (modeling labels) given that a random vector $\mathbf{X}$ (modeling feature vectors) equals $\mathbf{x}$. Let $L_{\mathrm{MAE}}(y|\mathbf{x}) := E[|Y - y| \mid \mathbf{X} = \mathbf{x}]$, i.e., the conditional expectation of the absolute-error loss given $\mathbf{X} = \mathbf{x}$. Express $h(\mathbf{x}) := \operatorname*{argmin}_{y \in \mathbb{R}} L_{\mathrm{MAE}}(y|\mathbf{x})$ in terms of $p(y|\mathbf{x})$.

Solution:

(a)

$$\boxed{\begin{aligned} h(\mathbf{x}) &= E\left[Y \mid \mathbf{X} = \mathbf{x})\right] = \text{(unique) } \textbf{mean} \text{ of the posterior distribution of the label given } \mathbf{x} \\ &= \sum_{y \in \mathcal{Y} \subseteq \mathbb{R}} y \cdot p(y|\mathbf{x}) \quad \text{(for discrete } Y\text{)} \\ &= \int_{y \in \mathbb{R}} y \cdot p(y|\mathbf{x})\, dy \quad \text{(for continuous } Y\text{)} \end{aligned}}$$

Note that the conditional expectation $E[\cdot \mid \mathbf{X} = \mathbf{x}]$ represents a discrete summation if Y is a discrete random variable and a continuous integral if Y is a continuous random variable. Sums and integrals are linear operations which can be interchanged with differentiation.

As a function of y, $L_{\text{MSE}}(y|\mathbf{x}) := E[(Y-y)^2 \mid \mathbf{X} = \mathbf{x}]$ is continuous and differentiable everywhere on the real line $\mathbb{R}$. The first derivative of $L_{\text{MSE}}(y|\mathbf{x})$ with respect to the variable y is given by

$$\frac{d}{dy} L_{\text{MSE}}(y|\mathbf{x}) = E\left[\frac{d}{dy}(y-Y)^2 \;\middle|\; \mathbf{X} = \mathbf{x}\right] = E\left[2(y-Y) \mid \mathbf{X} = \mathbf{x}\right] = 2\left(y - E\left[Y \mid \mathbf{X} = \mathbf{x}\right]\right)$$

which is strictly negative for all $y < E\left[Y \mid \mathbf{X} = \mathbf{x}\right]$, zero at $y = E\left[Y \mid \mathbf{X} = \mathbf{x}\right]$, and strictly positive for all $y > E\left[Y \mid \mathbf{X} = \mathbf{x}\right]$. It follows that $L_{\text{MSE}}(y|\mathbf{x})$ is strictly decreasing for all $y < E\left[Y \mid \mathbf{X} = \mathbf{x}\right]$, strictly increasing for all $y > E\left[Y \mid \mathbf{X} = \mathbf{x}\right]$, and minimized at the (unique) point $y = E\left[Y \mid \mathbf{X} = \mathbf{x}\right]$.

(b) $\boxed{h(\mathbf{x}) = \textbf{median} \text{ of } p(y|\mathbf{x})\text{, the posterior distribution of label given } \mathbf{x}}$

Note: the median of the posterior distribution of the label given $\mathbf{x}$ need not be unique. Now, $L_{\text{MAE}}(y|\mathbf{x}) := E[|Y-y| \mid \mathbf{X} = \mathbf{x}]$ as a function of y is continuous everywhere on $\mathbb{R}$ and differentiable everywhere except at points $\tilde{y}$ where Y has positive probability mass given $\mathbf{X} = \mathbf{x}$, i.e., $P(Y = \tilde{y}|\mathbf{X} = \mathbf{x}) > 0$. For points other than these, the first derivative of $L_{\text{MAE}}(y|\mathbf{x})$ with respect to the variable y is given by

$$\begin{aligned}\frac{d}{dy} L_{\text{MAE}}(y|\mathbf{x}) &= E\left[\frac{d}{dy}|y-Y| \;\middle|\; \mathbf{X} = \mathbf{x}\right] \\ &= E\left[\operatorname{sign}(y-Y) \;\middle|\; \mathbf{X} = \mathbf{x}\right] \\ &= \Pr\left(Y < y \;\middle|\; \mathbf{X} = \mathbf{x}\right) - \Pr\left(Y > y \;\middle|\; \mathbf{X} = \mathbf{x}\right)\end{aligned}$$

Let $\eta_{\text{label}}(\mathbf{x})$ denote a median value of the posterior distribution of the label given $\mathbf{X} = \mathbf{x}$. Reasoning as in Problem 2.2(c), we can conclude that $L_{\text{MAE}}(y|\mathbf{x})$ is non-increasing for all $y < \eta_{\text{label}}(\mathbf{x})$, non-decreasing for all $y > \eta_{\text{label}}(\mathbf{x})$, and constant for all median values of the posterior distribution of the label given $\mathbf{x}$. Therefore $L_{\text{MAE}}(y|\mathbf{x})$ is minimized at any median value of the posterior distribution of the label given $\mathbf{x}$.