Midterm Examination period 2020 – Week 8 – Semester A

MTH786: Machine learning with Python

You should attempt ALL questions. Marks available are shown next to the questions.

In completing this midterm assessment, you may use books, notes, and the Internet. You may use calculators and computers, but you must show your work for any calculations you do. You must not seek or obtain help from anyone else.

At the start of your work, please **copy out and sign** the following declaration:

> I declare that my submission is entirely my own, and I have not sought or obtained help from anyone else.

All work should be **handwritten** and should **include your student number**.

The exam is available for a period of **24 hours**. Upon accessing the exam, you will have **150 minutes** in which to complete and submit this assessment.

When you have finished your work:

- scan your work, convert it to a **single PDF file**, and submit this file on QMPlus;
- e-mail a copy to **maths@qmul.ac.uk** with your student number and the module code in the subject line;
- with your e-mail, include a photograph of the first page of your work together with either yourself or your student ID card.

You are expected to spend about **90 minutes** to complete the assessment, plus the time taken to scan and upload your work. Please try to upload your work well before the end of the submission window, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

Examiners: M. Benning

© Queen Mary University of London (2020)

Continue to next page

The notation log refers to the natural logarithm.

Question 1 [50 marks].

(a) For a uniform (and absolutely continuous) random variable X in $[0,1]$ compute the expectation of $f(X)$ for

$$f(x) := \begin{cases} a\log(x+b) & x \in [0,1/2] \\ 0 & \text{otherwise} \end{cases},$$

where a and b are the maximum of the seventh, respectively eighth, digit of your student ID and one. **[10]**

(b) Compute the probability $P(a \leq X \leq b)$ that an exponentially distributed random variable X with parameter $\lambda = 1/b$ lies in the interval $[a,b]$. Here a and b are the last two digits of your student ID that are ordered such that $a < b$. If $a = b$, consider a and $b+1$ instead. **[10]**

(c) Compute the gradient of the function $J(x) := \frac{1}{2}\langle Q(x-y),(x-y)\rangle$ for fixed $y \in \mathbb{R}^n$. Here $Q \in \mathbb{R}^{n\times n}$ is a (square) matrix. What does the gradient simplify to if Q is also symmetric? **[10]**

(d) Compute the MSE for the 1-parameter model by hand:

$$\mathrm{MSE}(w_0) = \frac{1}{2s}\sum_{i=1}^{s} |y_i - w_0|^2$$

Fill in the missing entries of the following table:

	$w_0 = -7$	$w_0 = -5$	$w_0 = -3$	$w_0 = -1$	$w_0 = 1$	$w_0 = 3$	$w_0 = 5$
y_1 = -1 y_2 = -2 y_3 = -3 y_4 = -4							
$\mathrm{MSE}(w)\cdot 2s$							
y_5 = 10 d							
$\mathrm{MSE}(w)\cdot 2s$							

Here d is the maximum of the seventh digit of your student ID and one. **[10]**

(e) Repeat the same exercise for what is known as the Mean Absolute Error (MAE), i.e.

$$\mathrm{MAE}(w_0) = \frac{1}{s}\sum_{i=1}^{s} |y_i - w_0|\,.$$

Fill in the missing entries of the following table:

	$w_0 = -7$	$w_0 = -5$	$w_0 = -3$	$w_0 = -1$	$w_0 = 1$	$w_0 = 3$	$w_0 = 5$
y_1 = -1 y_2 = -2 y_3 = -3 y_4 = -4							
$\mathrm{MAE}(w) \cdot s$							
$y_5 = 10\,d$							
$\mathrm{MAE}(w) \cdot s$							

The value d is once more the maximum of the seventh digit of your student ID and one. What do you observe, in particular with regards to the outlier y_5? **[10]**

Solution:

(a) The expectation for a uniform and absolutely continuous random variable X in $[0, 1]$ simply reads

$$\mathbb{E}_x[x] = \int_0^1 x \, dx \,.$$

Hence, we compute

$$\begin{aligned}\mathbb{E}_x[f(x)] &= \int_0^{\frac{1}{2}} a \log(x+b) \, dx = a \left((x+b) \log(x+b) - x \right)\Big|_0^{\frac{1}{2}} \\ &= a \left(\frac{1+2b}{2} \log\left(\frac{1+2b}{2} \right) - \frac{1}{2} \right) - a \left(b \log(b) \right) \\ &= a \left(\frac{1}{2} \log\left(1+2b\right) - \log(2) - \frac{1}{2} + b \log\left(1 + \frac{1}{2b}\right) \right) \,.\end{aligned}$$

If the student id digits are $a = 3$ and $b = 7$, we compute $\mathbb{E}_x[f(x)] \approx 2.971$ for example.

(b) The probability of a random variable X in the interval $[a, b]$ can be computed via

$$P(a \leq X \leq b) = \int_a^b \rho(x) \, dx \,.$$

In this exercise, ρ is the PDF of an exponential distribution with parameter $\lambda = 1/b$, i.e.

$$\rho(x) = \begin{cases} \frac{\exp\left(-\frac{x}{b}\right)}{b} & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \,.$$

Hence, we compute

$$P(a \leq X \leq b) = \frac{1}{b} \int_a^b e^{-\frac{x}{b}} \, dx = e^{-\frac{a}{b}} - \frac{1}{e} \,.$$

If the student id digits are $a = 2$ and $b = 5$ for example, we have $P(2 \leq X \leq 5) \approx 0.30244$.

(c) In order to compute the gradient, we need to compute the partial derivatives of J w.r.t. individual arguments x_l, for $l \in \{1, \ldots, n\}$, i.e.

$$\frac{\partial}{\partial x_l} J(x) = \frac{\partial}{\partial x_l} \frac{1}{2} \langle Q(x-y), x-y \rangle = \frac{\partial}{\partial x_l} \frac{1}{2} \sum_{i=1}^n \left(\sum_{j=1}^n Q_{ij}(x_j - y_j) \right) (x_i - y_i) \,.$$

With the product rule we easily compute

$$\frac{\partial}{\partial x_l} J(x) = \frac{1}{2} \left(\sum_{i=1}^n Q_{il}(x_i - y_i) + \sum_{j=1}^n Q_{lj}(x_j - y_j) \right) = \left(\frac{1}{2} \left(Q^\top + Q \right) (x-y) \right)_l \,.$$

Hence, the gradient ∇J reads $\nabla J(x) = \frac{1}{2} \left(Q^\top + Q \right) (x - y)$ for the column vector definition of the gradient, or $\nabla J(x) = \frac{1}{2}(x-y)^\top \left(Q^\top + Q \right)$ for the row-vector definition (both are acceptable solutions). The gradient simplifies to $\nabla J(x) = Q(x - y)$ if Q is symmetric, since $Q^\top = Q$ and thus $Q^\top + Q = 2Q$.

(d) The MSE for the 1-parameter model for $d = 1$ reads

	$w_0 = -7$	$w_0 = -5$	$w_0 = -3$	$w_0 = -1$	$w_0 = 1$	$w_0 = 3$	$w_0 = 5$
y_1 = -1	36	16	4	0	4	16	36
y_2 = -2	25	9	1	1	9	25	4
y_3 = -3	16	4	0	4	16	36	64
y_4 = -4	9	1	1	9	25	49	81
$\text{MSE}(w) \cdot 2s$	86	30	6	14	54	126	230
y_5 = 10	289	225	169	121	81	49	25
$\text{MSE}(w) \cdot 2s$	375	255	175	135	135	175	255

(e) The MAE for the 1-parameter model for $d = 1$ reads

	$w_0 = -7$	$w_0 = -5$	$w_0 = -3$	$w_0 = -1$	$w_0 = 1$	$w_0 = 3$	$w_0 = 5$
y_1 = -1	6	4	2	0	2	4	6
y_2 = -2	5	3	1	1	3	5	7
y_3 = -3	4	2	0	2	4	6	8
y_4 = -4	3	1	1	3	5	7	9
$\text{MAE}(w) \cdot s$	18	10	4	6	14	22	30
y_5 = 10	17	15	13	11	9	7	5
$\text{MAE}(w) \cdot s$	35	25	17	17	23	29	35

The MAE is less sensitive to extreme outliers compared to the MSE.

Question 2 [50 marks].

(a) Suppose we have s i.i.d. samples $x_1, \ldots, x_s$ that are drawn from a discrete Poisson distribution with parameter λ. Write down the likelihood for the data and use the maximum likelihood principle to compute the parameter λ. **[10]**

(b) Show that the function $f(x) = \alpha|x|$ is convex for fixed $\alpha > 0$. **[10]**

(c) Compute the proximal map for the function $f(x) = \frac{1}{2}\langle Qx, x\rangle$ for a square and symmetric matrix $Q \in \mathbb{R}^n$. **[10]**

(d) Show that the proximal map of the characteristic function

$$f(x) = \begin{cases} 0 & x \in [0,1] \\ \infty & x \notin [0,1] \end{cases}$$

is

$$(I + \partial f)^{-1}(z) = \begin{cases} 1 & z > 1 \\ z & z \in [0,1] \\ 0 & z < 0 \end{cases} .$$

[10]

(e) The rectifier or ramp function $f(x) = \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases}$ can also be written as the maximisation problem

$$f(x) = \max_{p \in [0,1]} xp .$$

Show that the smoothed ramp function

$$f_\tau(x) := \max_{p \in [0,1]} xp - \frac{\tau}{2}|p|^2$$

for a parameter $\tau > 0$ has the closed-form solution

$$f_\tau(x) = \begin{cases} x - \frac{\tau}{2} & x > \tau \\ \frac{1}{2\tau}x^2 & x \in [0,\tau] \\ 0 & x < 0 \end{cases} .$$

[10]

Solution:

(a) The likelihood for the data reads

$$\rho(x_1, \ldots, x_s|\lambda) = \prod_{i=1}^{s} \frac{\lambda^{x_i} \exp(-\lambda)}{x_i!} \,. \tag{1}$$

We use the maximum likelihood principle and minimise the negative logarithm of (1) w.r.t. λ, i.e.

$$\begin{aligned}
\hat{\lambda} &= \arg\min_{\lambda\in\mathbb{R}} -\log\left(\prod_{i=1}^{s} \frac{\lambda^{x_i}\exp(-\lambda)}{x_i!}\right) \\
&= \arg\min_{\lambda\in\mathbb{R}} -\sum_{i=1}^{s} \log\left(\frac{\lambda^{x_i}\exp(-\lambda)}{x_i!}\right) \\
&= \arg\min_{\lambda\in\mathbb{R}} \sum_{i=1}^{s} \left[\log(x_i!) - \log\left(\lambda^{x_i}\exp(-\lambda)\right)\right] \\
&= \arg\min_{\lambda\in\mathbb{R}} -\sum_{i=1}^{s} \left[\log\left(\lambda^{x_i}\right) + \log\left(\exp(-\lambda)\right)\right] \\
&= \arg\min_{\lambda\in\mathbb{R}} \sum_{i=1}^{s} \left[\lambda - x_i \log\left(\lambda\right)\right] \\
&= \arg\min_{\lambda\in\mathbb{R}} s\lambda - \sum_{i=1}^{s} x_i \log(\lambda) \,.
\end{aligned}$$

Computing the gradient of the function w.r.t. λ and setting it to zero yields

$$0 = s - \sum_{i=1}^{s} \frac{x_i}{\hat{\lambda}} \,.$$

Solving for $\hat{\lambda}$ then yields

$$\hat{\lambda} = \frac{1}{s}\sum_{i=1}^{s} x_i \,.$$

Hence, the parameter that maximises the likelihood is the mean of all samples $x_1, \ldots, x_s$.

(b) With the triangle inequality we instantly observe

$$\begin{aligned}
f(\lambda x + (1-\lambda)y) &= \alpha|\lambda x + (1-\lambda)y| \leq \alpha|\lambda x| + \alpha|(1-\lambda)y| \\
&\leq \lambda\alpha|x| + (1-\lambda)\alpha|y| = \lambda f(x) + (1-\lambda)f(y) \,,
\end{aligned}$$

for any $x, y \in \mathbb{R}$, $\alpha > 0$ and $\lambda \in [0,1]$. Hence, $f(x) = \alpha|x|$ is convex.

(c) The proximal map for the function $f(x) = \frac{1}{2}\langle Qx, x\rangle$ reads

$$(I + \partial f)^{-1}(z) = \arg\min_{x\in\mathbb{R}^n} \left\{ \frac{1}{2}\|x - z\|^2 + \frac{1}{2}\langle Qx, x\rangle \right\} .$$

The objective function is differentiable, so we can simply compute the gradient and set it to zero. From Question 1, Exercise (c), we already know that the gradient of $\frac{1}{2}\langle Qx, x\rangle$ for symmetric Q is Qx, hence, the full gradient set to zero reads

$$(I + \partial f)^{-1}(z) - z + Q(I + \partial f)(z) = 0 .$$

Solving this for $(I + \partial f)(z)$ yields

$$(I + \partial f)^{-1}(z) = (I + Q)^{-1}z .$$

(d) The proximal map is

$$\begin{aligned}(I + \partial f)^{-1}(z) &= \arg\min_{x} \left\{ \frac{1}{2}|x - z|^2 + f(x) \right\} \\ &= \arg\min_{x\in[0,1]} \left\{ \frac{1}{2}|x - z|^2 \right\} .\end{aligned}$$

We consider the three cases $z \in [0, 1]$, $z < 0$ and $z > 1$. If $z \in [0, 1]$, we easily observe that $x = z$ yields $\frac{1}{2}|x - z|^2 + f(x) = 0$, which is the global minimum since $\frac{1}{2}|x - z|^2 \geq 0$ and $f(z) \geq 0$ by definition. For $z < 0$ we observe that $x = 0$ makes the objective smallest. The last part can be verified through

$$\begin{aligned} & \frac{1}{2}| - z|^2 \leq \frac{1}{2}|x - z|^2 \\ \Leftrightarrow \quad & |z|^2 \leq |x - z|^2 \\ \Leftrightarrow \quad & |z|^2 \leq |x|^2 - 2xz + |z|^2 \\ \Leftrightarrow \quad & 2xz \leq |x|^2 \\ \Leftrightarrow \quad & 2z \leq x .\end{aligned}$$

The last inequality is always true, since we assumed $z < 0$ and $x \geq 0$ by definition. If $z > 1$ we observe that $x = 1$ makes the objective smallest, which we can verify in similar fashion as for the case $z < 0$. Hence, the proximal map reads

$$(I + \partial f)^{-1}(z) = \min\left(1, \max(0, z)\right) .$$

(e) Similar to Coursework 5, we can reformulate f as $f(x) = x\hat{p} - \tau\hat{p}^2/2$ with $\hat{p}$ satisfying

$$\begin{aligned}\hat{p} &= \arg\max_{p\in[0,1]} xp - \frac{\tau}{2}p^2 \\ &= \arg\min_{p\in[0,1]} \frac{\tau}{2}p^2 - xp \\ &= \arg\min_{p\in[0,1]} \frac{\tau}{2}\left(p - \frac{x}{\tau}\right)^2 .\end{aligned}$$

This is exactly the proximal map from Question 2 (e); hence, the solution is

$$\hat{p} = \begin{cases} 1 & x > \tau \\ \frac{x}{\tau} & x \in [0, \tau] \\ 0 & \text{otherwise} \end{cases} .$$

As a direct consequence, the function f reads

$$f(x) = \begin{cases} x - \frac{\tau}{2} & x > \tau \\ \frac{1}{2\tau}x^2 & x \in [0, \tau] \\ 0 & \text{otherwise} \end{cases} .$$

End of Paper.

© Queen Mary University of London (2020)