Queen Mary
University of London

Mid-term examination period 2021 – November – Semester A

MTH786P/U: Machine learning with Python

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

In completing this assessment:

- **You may use books and notes.**
- **You may use calculators and computers, but you must show your working for any calculations you do.**
- **You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.**
- **You must not seek or obtain 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:

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- 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.

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. Poplavskyi, N. Otter

 Continue to next page

Question 1 [50 marks]. Let $\sigma(w) = \frac{1}{1+\mathrm{e}^{-w}}$ be the logistic function. Let $\mathbf{x} = \{x_i\}_{i=1}^s$ be some data samples, where $x_i \in \mathbb{R}$ for all $i = 1, \ldots, s$ and

$$E(\mathbf{w} = (w_1, w_2, \ldots, w_s)) = -\log \sigma\left(\mathbf{x}^\top \mathbf{w}\right) = \log\left(1 + \mathrm{e}^{-\sum_{i=1}^{s} x_i w_i}\right),$$

be the logistic regression cost function.

(a) Prove $E(\mathbf{w})$ is a convex function.

Hint: You may wish use the fact that a twice differentiable function $\sigma(w) : \mathbb{R} \to \mathbb{R}$ is convex if and only if its second derivative satisfies

$$\sigma''(w) \geq 0, \quad \forall w \in \mathbb{R}.$$

Solution: We start by showing that $\ell(z) = -\log \sigma(z)$ is a convex function. We can prove this by evaluating the second derivative of $\ell(z)$. **[2]**

$$\frac{\mathrm{d}^2}{\mathrm{d}\, z^2}\ell(z) = \frac{\mathrm{d}^2}{\mathrm{d}\, z^2}\log\left(1+\mathrm{e}^{-z}\right) = -\frac{\mathrm{d}}{\mathrm{d}\, z}\frac{\mathrm{e}^{-z}}{1+\mathrm{e}^{-z}} = -\frac{\mathrm{d}}{\mathrm{d}\, z}\frac{1}{1+\mathrm{e}^{z}} = \frac{\mathrm{e}^{z}}{(1+\mathrm{e}^{z})} > 0. \quad [4]$$

The second derivative is positive and thus the function $\ell(z)$ is convex. **[2]**
The logistic cost function $L(\mathbf{w})$ can be written as a composition of the convex function $\ell(z)$ and a linear function $h(\mathbf{w}) = \mathbf{x}^\top \mathbf{w}$.
Therefore, as it was shown in the lecture, $L(\mathbf{w})$ is convex. **[2]**

(b) Show that $E(\mathbf{w})$ is an L-smooth function for $L = \frac{1}{4}\|\mathbf{x}\|^2$.

Hint: You may wish to use the following inequality (without proving it)

$$|\sigma(a) - \sigma(b)| \leq \frac{1}{4}|a - b|,$$

valid for any $a, b \in \mathbb{R}$.

Solution: Function $E(\mathbf{w}) : \mathbb{R}^s \to \mathbb{R}$ is L-smooth if for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^s$ one has

$$\|\nabla E(\mathbf{u}) - \nabla E(\mathbf{v})\| \leq L\|\mathbf{u} - \mathbf{v}\|. \quad [2]$$

The gradient of $E(\mathbf{w})$ is given by

$$\nabla E(\mathbf{w}) = \left(\frac{\partial}{\partial w_1}E(\mathbf{w}), \ldots, \frac{\partial}{\partial w_s}E(\mathbf{w}),\right). \quad [1]$$

Corresponding partial derivatives are equal to

$$\frac{\partial}{\partial w_j}E(\mathbf{w}) = -x_j\frac{\mathrm{e}^{-\sum_{i=1}^{s} x_i w_i}}{1+\mathrm{e}^{-\sum_{i=1}^{s} x_i w_i}} = -x_j \cdot \sigma\left(-\mathbf{x}^\top \mathbf{w}\right), \quad [2]$$

 Continue to next page

and the gradient is thus equal to

$$\nabla E(\mathbf{w}) = -\mathbf{x}\sigma\left(-\mathbf{x}^\top \mathbf{w}\right). \quad [\mathbf{1}]$$

Plugging in the above we get **[4]**

$$\begin{aligned} \|\nabla E(\mathbf{u}) - \nabla E(\mathbf{v})\| &= \left\|\mathbf{x}\sigma\left(-\mathbf{x}^\top \mathbf{v}\right) - \mathbf{x}\sigma\left(-\mathbf{x}^\top \mathbf{u}\right)\right\| \\ &= \|\mathbf{x}\| \left|\sigma\left(-\mathbf{x}^\top \mathbf{v}\right) - \sigma\left(-\mathbf{x}^\top \mathbf{u}\right)\right| \\ &\overset{(1)}{\leq} \frac{1}{4}\|\mathbf{x}\| \left|\mathbf{x}^\top(\mathbf{u}-\mathbf{v})\right| \leq \frac{1}{4}\|\mathbf{x}\|^2 \|\mathbf{u}-\mathbf{v}\|, \end{aligned}$$

where in (1) we have used an inequality

$$|\sigma(a) - \sigma(b)| \leq \frac{1}{4}|a-b|,$$

for $a = -\mathbf{x}^\top \mathbf{u}$ and $b = -\mathbf{x}^\top \mathbf{v}$.

We say that a random variable ξ follows the Rayleigh distribution with parameter α if its probability density function is given by

$$p(x|\alpha) = \begin{cases} 2\alpha x \mathrm{e}^{-\alpha x^2}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

Let the parameter α be equal to the maximum between the fifth digit of your student ID and one.

(c) Calculate the variance $\mathrm{Var}[\xi]$ of a random variable ξ drawn from the Rayleigh distribution with the parameter α as described above.

Solution: The variance of a random variable is given by

$$\mathrm{Var}[\xi] = \mathbb{E}\left[\xi^2\right] - \mathbb{E}^2[\xi], \quad [\mathbf{1}]$$

where

$$\mathbb{E}[\xi] = \int x p(x|\alpha)\,\mathrm{d}x, \quad \mathbb{E}\left[\xi^2\right] = \int x^2 p(x|\alpha)\,\mathrm{d}x. \quad [\mathbf{2}]$$

The mean is then equal to **[3]**

$$\begin{aligned} \mathbb{E}[\xi] &= \int_0^\infty 2\alpha x^2 \mathrm{e}^{-\alpha x^2}\,\mathrm{d}x = -\int_0^\infty x\,\mathrm{d}\mathrm{e}^{-\alpha x^2} = \int_0^\infty \mathrm{e}^{-\alpha x^2}\,\mathrm{d}x = \left| x \to s/\sqrt{2\alpha} \right| \\ &= \frac{1}{\sqrt{2\alpha}} \int_0^\infty \mathrm{e}^{-\frac{s^2}{2}}\,\mathrm{d}s = \frac{1}{2}\sqrt{\frac{\pi}{\alpha}}. \end{aligned}$$

The second moment is equal to **[3]**

$$\begin{aligned} \mathbb{E}[\xi] &= \int_0^\infty 2\alpha x^3 \mathrm{e}^{-\alpha x^2}\,\mathrm{d}x = -\int_0^\infty x^2\,\mathrm{d}\mathrm{e}^{-\alpha x^2} = \int_0^\infty 2x\mathrm{e}^{-\alpha x^2}\,\mathrm{d}x \\ &= \frac{1}{\alpha}\int_0^\infty \mathrm{d}\mathrm{e}^{-\alpha x^2} = \frac{1}{\alpha}. \end{aligned}$$

 Continue to next page

And the variance is thus equal to

$$\mathrm{Var}\left[\xi\right] = \frac{1}{\alpha} - \frac{\pi}{4\alpha} = \frac{4-\pi}{4\alpha}. \qquad [\mathbf{1}]$$

(d) Suppose you get s i.i.d. samples $\{\xi_i\}_{i=1}^{s}$ drawn from the Rayleigh distribution with the parameter α described above. Derive the optimisation problem for the negative log-likelihood estimator for α of the form

$$\hat{\alpha} = \min_{\alpha>0} E\left(\alpha\right),$$

where $E\left(\alpha\right)$ is an energy function.

Solution: Random samples ξ_i are i.i.d, and thus the likelihood of getting corresponding values is equal to the product of independent likelihoods for each data sample. **[2]**

$$p\left(\xi_1, \xi_2, \ldots, \xi_s|\alpha\right) = \prod_{i=1}^{s} p\left(\xi_i|\alpha\right) = \prod_{i=1}^{s} 2\alpha\xi_i \mathrm{e}^{-\alpha\xi_i^2}. \qquad [\mathbf{3}]$$

The negative log-likelihood is equal to

$$-\log p\left(\xi_1, \ldots, \xi_s|\alpha\right) = \sum_{i=1}^{s} \alpha\xi_i^2 - \log\alpha - \log\left(2\xi_i\right) = \alpha\sum_{i=1}^{s}\xi_i^2 - s\log\alpha - \sum_{i=1}^{s}\log\left(2\xi_i\right). \qquad [\mathbf{3}]$$

Maximisation of the likelihood is equivalent to a minimisation of negative log-likelihood. We also note the last term in the above expression doesn't depend on α and thus the optimisation problem can be written as

$$\hat{\alpha} = \arg\min_{\alpha>0}\left[\alpha\sum_{i=1}^{s}\xi_i^2 - s\log\alpha\right]. \qquad [\mathbf{2}]$$

(e) Solve the maximum likelihood estimator problem for the parameter α based on these samples.

Solution: Let

$$E\left(\alpha\right) = \alpha\sum_{i=1}^{s}\xi_i^2 - s\log\alpha,$$

be a cost function from the last question. The energy function E is equal to a sum of linear functions and a convex one, thus it is convex and its minimiser can be found by solving **[2]**

$$\nabla E\left(\hat{\alpha}\right) = 0.$$

The gradient (in this case just a derivative) of the energy function is equal to

$$E'\left(\alpha\right) = \sum_{i=1}^{s}\xi_i^2 - \frac{s}{\alpha}. \qquad [\mathbf{4}]$$

The solution of the optimisation problem is then equal to

$$\sum_{i=1}^{s} \xi_i^2 - \frac{s}{\hat{\alpha}} = 0 \Rightarrow \hat{\alpha} = \frac{s}{\sum\limits_{i=1}^{s} \xi_i^2}. \qquad [4]$$

Question 2 [50 marks]. Suppose you are given $s = 5$ samples $\{(x_i, y_i)\}_{i=1}^{5}$ with

$$\begin{aligned} (x_1, y_1) &= (-2, -1), \\ (x_2, y_2) &= (-1, 1), \\ (x_3, y_3) &= (0, 3), \\ (x_4, y_4) &= (1, 5), \\ (x_5, y_5) &= (2, 7). \end{aligned}$$

In this question you are asked to build a one-feature linear regression model

$$y_i \approx f_\theta(x_i) := \theta x_i,$$

where θ is the only weight parameter.

 Continue to next page

(a) Start with the mean-squared error (MSE) as a measure of a deviation, where

$$\mathrm{MSE}(\theta) = \frac{1}{2s}\sum_{i=1}^{s}\left(y_i - f_\theta(x_i)\right)^2. \tag{1}$$

Run the grid search algorithm to minimise $\mathrm{MSE}(\theta)$ over the grid $\theta \in G$, with $G = \{1, 2, 3, 4\}$. Fill in the missing entries of the following table by evaluating $y_i - \theta x_i$ for corresponding values of the parameters, and find the minimiser $\hat{\theta}$ given by

$$\hat{\theta} = \arg\min_{\theta \in G} \mathrm{MSE}(\theta).$$

	$\theta = 1$	$\theta = 2$	$\theta = 3$	$\theta = 4$
$x_1 = -2,\ y_1 = -1$				
$x_2 = -1,\ y_2 = 1$				
$x_3 = 0,\ y_3 = 3$				
$x_4 = 1,\ y_4 = 5$				
$x_5 = 2,\ y_5 = 7$				
$\mathrm{MSE}(\theta)$				

Solution: The table reads **[8]**

	$\theta = 1$	$\theta = 2$	$\theta = 3$	$\theta = 4$
$x_1 = -2,\ y_1 = -1$	1	3	5	7
$x_2 = -1,\ y_2 = 1$	2	3	4	5
$x_3 = 0,\ y_3 = 3$	3	3	3	3
$x_4 = 1,\ y_4 = 5$	4	3	2	1
$x_4 = 2,\ y_4 = 7$	5	3	1	-1
$\mathrm{MSE}(\theta)$	5.5	4.5	5.5	8.5

The minimiser is then equal to $\hat{\theta} = 2$. **[2]**

(b) Derive the gradient descent update formula that aims at minimising $\mathrm{MSE}(\theta)$.

Solution: The gradient descent update rule takes the form

$$\theta^{(k+1)} = \theta^{(k)} - \tau \nabla \mathrm{MSE}\left(\theta^{(k)}\right). \tag{2}$$

The gradient of MSE defined in (1) is equal to

$$\nabla \mathrm{MSE}(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\frac{1}{2s}\sum_{i=1}^{s}\left(y_i - \theta x_i\right)^2 = \frac{1}{s}\sum_{i=1}^{s} x_i\left(\theta x_i - y_i\right) = \theta\overline{\mathbf{x}^2} - \overline{\mathbf{xy}}, \tag{6}$$

where $\overline{\mathbf{x}^2} = \frac{1}{s}\sum_{i=1}^{s} x_i^2$ and $\overline{\mathbf{xy}} = \frac{1}{s}\sum_{i=1}^{s} x_i y_i$. The gradient descent update formula then takes the form

$$\theta^{(k+1)} = \theta^{(k)}\left(1 - \tau\overline{\mathbf{x}^2}\right) + \tau\overline{\mathbf{xy}}. \tag{2}$$

(c) Let $\theta^{(0)}$ be the sixth digit of your student ID. Perform two steps of the gradient descent method with the step-size set to $\tau = \frac{1}{2}$ to evaluate $\theta^{(1)}$ and $\theta^{(2)}$. Comment on your findings.

Solution: For given data samples one can easily evaluate

$$\overline{\mathbf{x}^2} = \frac{1}{5}(4+1+0+1+4) = 2, \quad \overline{\mathbf{xy}} = \frac{1}{5}(2-1+0+5+14) = 4. \qquad [\mathbf{2}]$$

Plugging the above to the gradient descent update rule one gets

$$\theta^{(1)} = \theta^{(0)}\left(1 - \frac{1}{2}\cdot 2\right)\theta^{(0)} + \frac{1}{2}\cdot 4 = 0 + 2 = 2. \qquad [\mathbf{2}]$$

$$\theta^{(2)} = \theta^{(0)}\left(1 - \frac{1}{2}\cdot 2\right)\theta^{(1)} + \frac{1}{2}\cdot 4 = 0 + 2 = 2. \qquad [\mathbf{2}]$$

Because the sequence has stabilised, it is clear that the gradient descent algorithm achieved a minimiser. Thus the minimiser for an MSE function defined in (1) is equal to $\hat{\theta} = 2$. **[4]**

(d) Let a and b be the maximum between the seventh, respectively eighth, digit of your student ID and one. The leaky rectifier function

$$R_{a,b}(x) = \begin{cases} ax, & x \geq 0, \\ -bx, & x < 0. \end{cases}$$

can also be written via the maximisation problem

$$R_{a,b}(x) = \max_{p\in[-b,a]} xp\,.$$

Show that the smoothed leaky rectifier function

$$R_{a,b,\tau}(x) := \max_{p\in[-b,a]} xp - \frac{\tau}{2}|p|^2$$

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

$$R_{a,b,\tau}(x) = \begin{cases} ax - \frac{\tau}{2}a^2, & x > a\tau, \\ \frac{x^2}{2\tau}, & -b\tau < x \leq a\tau, \\ -bx - \frac{\tau}{2}b^2, & x \leq -b\tau. \end{cases}$$

Solution: Let $f_x(p) = xp - \frac{\tau}{2}p^2$. Then the smoothed parametric rectifier function $R_{a,b,\tau}$ is equal to

$$R_{a,b,\tau}(x) = \max_{p\in[-b,a]} f_x(p)\,.$$

Let us first evaluate the derivative of $f_x(p)$:

$$f_x'(p) = x - \tau p. \qquad [\mathbf{1}]$$

Let us now consider three cases:

- $x > a\tau$: then $f'_x(p) > a\tau - \tau p \geq 0$, for every $p \in [-b, a]$. This yields that the function $f_x(p)$ is an increasing function over the interval $[-b, a]$ and

$$\max_{p\in[-b,a]} f_x(p) = f_x(a),$$

and $R_{a,b,\tau}(x) = ax - \frac{\tau}{2}a^2$. **[3]**

- $-b\tau < x \leq a\tau$: then $f_x(p)$ is a quadratic function whose maximum is achieved at the point p^* such that $f'_x(p^*) = 0$. It is easy to see that $p^* = x/\tau$ and thus $R_{a,b,\tau}(x) = p^*x - \frac{\tau}{2}(p^*)^2 = \frac{1}{2\tau}x^2$. **[3]**

- $x \leq -b\tau$: then $f'_x(p) \leq -b\tau - \tau p \leq 0$, for every $p \in [-b, a]$. This yields that the function $f_x(p)$ is a decreasing function over the interval $[-b, a]$ and

$$\max_{p\in[-b,a]} f_x(p) = f_x(-b),$$

and $R_{a,b,\tau}(x) = -bx - \frac{\tau}{2}b^2$. **[3]**

(e) Consider a regularised optimisation problem of the form

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}} \left[\mathrm{MSE}(\theta) + R_{a,b,\tau}(\theta)\right], \tag{2}$$

where the MSE and the parametric rectifier $R_{a,b,\tau}$ functions are defined above. Derive the gradient descent update formula that aims at solving (2). Let $\theta^{(0)}$ be the seventh digit of your student ID. Perform two steps of the gradient descent starting from $\theta^{(0)}$ with step-size set to $\tau = \frac{1}{2}$ to evaluate $\theta^{(1)}$ and $\theta^{(2)}$.

Solution: The gradient descent update rule takes the form

$$\theta^{(k+1)} = \theta^{(k)} - \tau\nabla\left[\mathrm{MSE} + R_{a,b,\tau}\right]\left(\theta^{(k)}\right). \quad [\mathbf{1}]$$

The gradient of a regularised MSE defined in (2) is equal to

$$\nabla\left[\mathrm{MSE}(\theta) + R_{a,b,\tau}(\theta)\right] = \theta\overline{\mathbf{x}^2} - \overline{\mathbf{xy}} + R'_{a,b,\tau}(\theta), \quad [\mathbf{1}]$$

where the derivative $R'_{a,b,\tau}(\theta)$ can be calculated as

$$R'_{a,b,\tau}(\theta) = \begin{cases} a, & x > a\tau, \\ \frac{x}{\tau}, & -b\tau < x \leq a\tau, \\ -b, & x \leq -b\tau. \end{cases} \quad [\mathbf{2}]$$

Combining the above one can write the gradient descent update rule as

$$\theta^{(k+1)} = \left(1 - \tau\overline{\mathbf{x}^2}\right)\theta^{(k)} + \tau\overline{\mathbf{xy}} - \begin{cases} a\tau, & \theta^{(k)} > a\tau, \\ \theta^{(k)}, & -b\tau < \theta^{(k)} \leq a\tau, \\ -b\tau, & \theta^{(k)} \leq -b\tau. \end{cases} \quad [\mathbf{2}]$$

 Continue to next page

Using the values obtained before we get

$$\theta^{(k+1)} = 2 - \begin{cases} \frac{a}{2}, & \theta^{(k)} > \frac{a}{2}, \\ \theta^{(k)}, & -\frac{b}{2} < \theta^{(k)} \leq \frac{a}{2}, \\ -\frac{b}{2}, & \theta^{(k)} \leq -\frac{b}{2}. \end{cases} \quad [\mathbf{2}]$$

The result of gradient descent execution depends on values a, b that are individual for every student. **[2]**

End of Paper.