Queen Mary
University of London

Late-Summer Examination period 2019

MTH786P: Machine Learning with Python

Duration: 3 hours

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

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

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any unauthorised notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: M. Benning

The terms MSE and MAE are abbreviations for Mean Squared Error and Mean Absolute Error, respectively. Note that for all multiple choice questions there will only be **one** correct answer.

Question 1. [34 marks]

(a) Which of the following properties is **not** true for a Hessian matrix?

A. The Hessian is symmetric.

B. The Hessian is always invertible.

C. The Hessian is the transpose of the Jacobi matrix of the gradient.

D. The Hessian is used in the Newton-Raphson method.

[5]

(b) Mini-batch gradient descent coincides with

A. stochastic gradient descent if every batch contains exactly one randomly chosen index

B. gradient descent if every batch contains exactly one index

C. the Newton-Raphson method if every batch contains all indices.

D. none of the above.

[5]

(c) 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:

	$w_0 = -3$	$w_0 = -2$	$w_0 = -1$	$w_0 = 0$	$w_0 = 1$	$w_0 = 2$	$w_0 = 3$
y_1 = -3 y_2 = -2 y_3 = -1 y_4 = 0 y_5 = 1							
$2\,\mathrm{MSE}(w_0)s$							
y_6 = -20							
$2\,\mathrm{MSE}(w_0)s$							

Some help: $23^2 = 529, 22^2 = 484, 21^2 = 441, 20^2 = 400, 19^2 = 361, 18^2 = 324, 17^2 = 289$. **[5]**

(d) Repeat the same exercise with the 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:

	$w_0 = -3$	$w_0 = -2$	$w_0 = -1$	$w_0 = 0$	$w_0 = 1$	$w_0 = 2$	$w_0 = 3$
y_1 = -3							
y_2 = -2							
y_3 = -1							
y_4 = 0							
y_5 = 1							
MAE(w_0)s							
y_6 = -20							
MAE(w_0)s							

What do you observe, in particular with regards to the outlier y_6? **[6]**

(e) Compute the gradient ∇L of the cost function $L : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$L(z_1, z_2) = z_1 \log\left(\frac{z_1}{z_2}\right) + z_2 - z_1 \,.$$

[5]

(f) Derive the gradient descent update formula that aims at minimising L as defined in (e). **[4]**

(g) Show that the Hessian $H_L(z_1, z_2)$ of L with respect to arguments z_1 and z_2 equals

$$H_L(z_1, z_2) = \frac{1}{z_2} \begin{pmatrix} \frac{z_2}{z_1} & -1 \\ -1 & \frac{z_1}{z_2} \end{pmatrix} .$$

Further verify that the Hessian is is positive semi-definite, i.e.

$$\begin{pmatrix} x & y \end{pmatrix} H_L(z_1, z_2) \begin{pmatrix} x \\ y \end{pmatrix} \geq 0$$

holds true for all $x, y \in \mathbb{R}$ and all $z_1 > 0$ and $z_2 > 0$. **[4]**

Question 2. [33 marks]

(a) Computing the gradient of a function and finding parameters that map the gradient to zero implies that

A. those parameters minimise the function globally.

B. those parameters maximise the function if the function is also concave.

C. those parameters maximise the function locally.

D. the function is not concave.

[4]

(b) Proximal gradient descent is a special case of

A. the Newton-Raphson method.

B. the grid search method.

C. gradient descent if the function R in the proximal mapping is chosen to be zero for all arguments.

D. stochastic gradient descent with batch-size one.

[4]

(c) For the function $f(x, y) = \sin(x + y)$ we observe

A. $\min_x \max_y f(x, y) < \max_y \min_x f(x, y)$.

B. $\min_x \max_y f(x, y) = \max_y \min_x f(x, y)$.

C. $\max_x \min_y f(x, y) < \min_y \max_x f(x, y)$.

D. $\max_x \min_y f(x, y) > \min_y \max_x f(x, y)$.

[2]

(d) State the definition of convexity of a function. **[4]**

(e) State the definition of the Euclidean vector norm (also known as two-norm). **[4]**

(f) State the definition of the matrix product Xw for a matrix $X \in \mathbb{R}^{s\times n}$ and a vector $w \in \mathbb{R}^n$ and show that it is linear, i.e.

$$X(aw + b) = aXw + Xb\,,$$

for $a \in \mathbb{R}$ and $b \in \mathbb{R}^n$. **[6]**

(g) Show that for given $X \in \mathbb{R}^{s\times n}$ and $y \in \mathbb{R}^s$ the affine-linear function

$$h(w) := Xw - y$$

satisfies $h(\lambda w + (1 - \lambda)v) = \lambda h(w) + (1 - \lambda)h(v)$ for all $w, v \in \mathbb{R}^n$ and $\lambda \in [0, 1]$. **[4]**

(h) Show that the MSE function of the form

$$\text{MSE}(w) = \frac{1}{2s}\|Xw - y\|^2$$

is convex.
Hint: use (g) to show that the composition $g(h(w))$ of a convex function g and an affine-linear function h is convex for all arguments $w \in \mathbb{R}^n$. **[5]**

Question 3. [33 marks]

(a) The problem to decide whether given fruits are either apples, bananas or strawberries is a

A. binary classification problem.

B. regression problem with continuous output variables.

C. problem that has no solution.

D. multi-class classification problem with more than two classes.

[6]

(b) Which of the following statements is generally **not** true for matrix factorisations?

A. They factorise matrices A into products of the form $A = BC$ for matrices B and C.

B. There is always a unique solution to the matrix factorisation problem.

C. They can be used to model recommender systems.

D. Singular value decomposition is a special case of matrix factorisation.

[6]

(c) Suppose $\{p_i\}_{i=1}^s$ with $\sum_{i=1}^s p_i = 1$ and $p_i \geq 0$. Show that for

$$L(x) = \sum_{i=1}^{s} p_i \ell_i(x)$$

the gradient and the expected value of the stochastic gradient coincide, i.e.

$$\mathbb{E}_{i\sim D}\left[\nabla \ell_i(x)\right] = \nabla L(x)\,,$$

where the expected value is defined with respect to $\{p_i\}_{i=1}^s$, i.e. $\mathbb{E}_{i\sim D}[y_i] = \sum_{i=1}^s p_i y_i$. **[6]**

(d) What is the maximum likelihood estimator (MLE)? Derive the MLE of a linear model assuming that the data $\{y_i\}_{i=1}^s$ are i.i.d. random variables with probability density function

$$p(y_i|x_i, w) = \frac{1}{\langle x_i, w\rangle} e^{-\frac{y_i}{\langle x_i, w\rangle}}\,.$$

- Derive $p(y|X, w)$ for $y = \begin{pmatrix} y_1 \\ \vdots \\ y_s \end{pmatrix} \in \mathbb{R}^s$, $x_i = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$, $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_s^T \end{pmatrix} \in \mathbb{R}^{s\times n}$.
- Compute the negative log-likelihood.
- Substitute $z_i = \log(\langle x_i, w\rangle)$ in the negative log-likelihood and simplify as much as possible.

[15]

End of Paper.