Queen Mary
University of London

Main Examination period 2020

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

© Queen Mary University of London (2020) **Turn Over**

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. The notation $\binom{n}{k}$ denotes the binomial coefficient $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

Question 1. [50 marks]

(a) Which one of the following strategies is **not** a regularisation strategy?

A. Adding a positive multiple of the squared two-norm of the weights to the MSE.

B. Early stopping of gradient descent.

C. Subtracting a constant from the MSE.

D. Adding a positive multiple of the one-norm of the weights to the MSE.

[5]

(b) The difference between incremental and stochastic gradient descent (with batch-size one) is

A. one is a descent method while the other one is not.

B. that for stochastic gradient descent the indices are drawn randomly whilst for incremental gradient descent the ordering is deterministic.

C. that one uses the Hessian in order to accelerate convergence.

D. only in name; otherwise they are identical.

[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:

	$w_0 = -5$	$w_0 = -3$	$w_0 = -1$	$w_0 = 0$	$w_0 = 1$	$w_0 = 3$	$w_0 = 5$
y_1 = -3 y_2 = -2 y_3 = -1 y_4 = 0 y_5 = 1							
2 MSE$(w_0)s$							
y_6 = -15							
2 MSE$(w_0)s$							

[7]

© Queen Mary University of London (2020)

(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:

	$w_0 = -5$	$w_0 = -3$	$w_0 = -1$	$w_0 = 0$	$w_0 = 1$	$w_0 = 3$	$w_0 = 5$
y_1 = -3							
y_2 = -2							
y_3 = -1							
y_4 = 0							
y_5 = 1							
$\mathrm{MAE}(w_0)s$							
y_6 = 15							
$\mathrm{MAE}(w_0)s$							

What do you observe, in particular with regards to the outlier y_6? **[7]**

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

$$L(x, y) = \frac{x}{y} - 1 - \log\left(\frac{x}{y}\right) .$$

[5]

(f) Show that L is scalar-invariant, i.e. $L(x, y) = L(c\,x, c\,y)$ for any scalar $c > 0$ and all arguments $x > 0, y > 0$. **[5]**

(g) Derive the corresponding gradient descent update formula for L as defined in (e). **[5]**

(h) Verify that the function $f_y(x) := L(x, y)$ is convex for fixed $y > 0$. **[5]**

(i) Show that the Hessian $H_L(x, y)$ of L with respect to arguments $x > 0$ and $y > 0$ equals

$$H_L(x, y) = \frac{1}{y^2}\begin{pmatrix} \frac{y^2}{x^2} & -1 \\ -1 & \frac{2x-y}{y} \end{pmatrix} .$$

Further verify that the Hessian is **not** positive semi-definite, i.e. there exist $x > 0, y > 0, a > 0$ and $b > 0$ such that

$$\begin{pmatrix} a & b \end{pmatrix} H_L(x, y) \begin{pmatrix} a \\ b \end{pmatrix} < 0 \,.$$

[6]

Question 2. [30 marks]

(a) When a function is convex and a minimiser exists, then this minimiser

A. is always unique.

B. is never unique.

C. is a global minimiser.

D. is a local but not a global minimiser.

[5]

(b) Proximal gradient descent applied to the lasso problem is known as

A. incremental soft-thresholding algorithm.

B. iterative soft-thresholding algorithm.

C. inertial soft-thresholding algorithm.

D. inertial soft-tissue algorithm.

[5]

(c) State the definitions of convexity and concavity of a function. **[4]**

(d) Verify that the function $f : \mathbb{R} \to \overline{\mathbb{R}}$, with

$$f(x) := \begin{cases} 0 & x \geq 0 \\ +\infty & x < 0 \end{cases}, \tag{1}$$

is convex. **[4]**

(e) State the definition of the proximal mapping $\text{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ of a convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$. **[4]**

(f) Derive a closed-form solution of the proximal mapping $\text{prox}_f : \mathbb{R} \to \mathbb{R}$ for the function defined in (1).
Hint: use case-differentiation to derive the proximal mapping. **[4]**

(g) Show that any linear function $f : \mathbb{R}^n \to \mathbb{R}$ is both convex and concave. **[4]**

© Queen Mary University of London (2020)

Question 3. [20 marks]

(a) The problem of estimating the weight of a person given their height based on other pairs of weight/height samples is

A. a multi-class classification problem with more than two classes.

B. a binary classification problem.

C. an unsupervised learning task.

D. a regression problem with continuous output.

[5]

(b) 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. samples of a binomial distributed random variable with probability mass function

$$p(y_i|x_i, w) = \binom{c}{y_i} \sigma(\langle x_i, w\rangle)^{y_i}(1 - \sigma(\langle x_i, w\rangle)^{c-y_i},$$

for a constant $c \in \mathbb{N}$ and a function $\sigma : \mathbb{R} \to [0, 1]$.

- 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.
- Insert $\sigma(x) = 1/(1 + e^{-x})$ in the negative log-likelihood and simplify as much as possible.

[15]

End of Paper.

© Queen Mary University of London (2020)