Queen Mary
University of London

Main 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. The notation $i \sim U$ suggests that i is drawn from a (discrete) uniform distribution.

Question 1. [34 marks]

(a) How can we prevent overfitting in polynomial regression?

A. Increase the degree of the polynomial.

B. Increase the number of data samples.

C. Decrease the number of data sampes.

D. Use gradient descent to fit the polynomial.

[5]

(b) If we try to solve $Ax = b$ for $x \in \mathbb{R}^n$ where $b \in \mathbb{R}^m$ is a vector and $A \in \mathbb{R}^{m \times n}$ an ill-conditioned matrix, then

A. small errors in b have almost no effect on the solution x.

B. large errors in b have almost no effect on the solution x.

C. small errors in b can lead to large errors in the solution x.

D. this means that the condition number of A is very small.

[5]

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

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

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 = -2 y_2 = -1 y_3 = 0 y_4 = 1 y_5 = 2							
2 MSE$(w_0)s$							
y_6 = 20							
2 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.

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

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

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

(e) For given data $y \in \{0, 1\}$ compute the gradient ∇L of the cost function

$$L(z_1, z_2) = \log\left(\exp(z_1) + \exp(z_2)\right) - \left(y\, z_1 + (1 - y)\, z_2\right) \,.$$

[5]

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

(g) Is the Newton-Raphson method, which aims at finding the root $\hat{z}$ of $\nabla L(\hat{z}) = 0$, well defined for L in (e)? If yes, please state the update formula. If not, please explain the problem. **[3]**

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 minimise the function if the function is also convex.

C. those parameters minimise the function locally.

D. the function is not convex.

[4]

(b) Projected gradient descent is a special case of

A. the grid search method.

B. the Newton-Raphson method.

C. stochastic gradient descent with batch-size one.

D. proximal gradient descent.

[4]

(c) The max-min inequality states $\max_x \min_y f(x,y) \leq \min_y \max_x f(x,y)$. In which of the following cases can we guarantee equality?

A. f is concave in its first and concave in its second argument.

B. f is convex in its first argument and non-convex in its second argument.

C. f is non-convex in its first argument and concave in its second argument.

D. f is convex in its first and concave in its second argument.

[4]

(d) State the definitions of convexity and concavity of a function. **[6]**

(e) State the definition of the inner product $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ and show that it is bilinear, i.e.

$$\langle ax + b, y \rangle = a\langle x, y \rangle + \langle b, y \rangle ,$$

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

(f) Show that for fixed $x \in \mathbb{R}^n$ and $y \in \mathbb{R}$ the affine-linear function

$$h(w) := \langle x, w \rangle - y$$

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]**

(g) Show that the MSE function of the form

$$\mathrm{MSE}(w) = \frac{1}{2s} \sum_{i=1}^{s} |\langle w, x_i \rangle - y_i|^2$$

is convex.
Hint: use (f) 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 a flight is delayed by more than four hours or not is a

A. regression problem with continuous output variables.
B. multi-class classification problem with more than two classes.
C. problem that has no solution.
D. binary classification problem.

[6]

(b) Which of the following statements is generally **not** true for singular value decomposition?

A. It splits a matrix A into a sum $A = B + C$.
B. It decomposes a matrix A into $A = U\Sigma V^T$.
C. It can be used to compress signals such as digital images.
D. It can be used for matrix factorisation.

[6]

(c) Show that for

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

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

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

[6]

(d) Derive the maximum likelihood estimator of a linear model assuming that the error follows a Laplacian distribution with location zero and scale 1, i.e.

$$y_i = \langle w, x_i\rangle + \varepsilon_i\,,$$

where each ε_i is an i.i.d. random variable that has a Laplace(0, 1) distribution, i.e. its probability density function is

$$p(\varepsilon_i) = \frac{1}{2}e^{-|\varepsilon_i|}$$

for all $i \in \{1, \ldots, s\}$.

- 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.
- Compare the negative log-likelihood to the MAE.

[15]

End of Paper.