UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2017/2018

ADVANCED MACHINE LEARNING

Duration: 120 mins

You must enter your Student ID and your ISS login ID (as a cross-check) on this page. You must not write your name anywhere on the paper.

Student ID:	
ISS ID:	

Question	Marks
A1	
B1	
B2	
B3	
Total	

Answer all parts of the question in section A (30 marks) and TWO questions from section B (35 marks each)

This examination is worth 60%. The coursework was worth 40%.

University approved calculators MAY be used.

A foreign language translation dictionary (paper version) is permitted provided it contains no notes, additions or annotations.

Each answer must be completely contained within the box under the corresponding question. No credit will be given for answers presented elsewhere.

You are advised to write using a soft pencil so that you may readily correct mistakes with an eraser.

You may use a blue book for scratch—it will be discarded without being looked at.

Section A

Question A 1

(a) Briefly describe the type of data where the following learning machines excel: (i) SVMs, (ii) Gradient Boosting and (iii) CNNs. *(6 marks)*

i

ii

iii

$\overline{6}$

(b) Show that for the mapping

$$\boldsymbol{x} = (x_1, x_2, x_3) \to \vec{\phi}(\boldsymbol{x}) = (x_1^2, x_2^2, x_3^2, \sqrt{2}\, x_1\, x_2,\ \sqrt{2}\, x_1\, x_3,\ \sqrt{2}\, x_2\, x_3)$$

the kernel $K(\boldsymbol{x}, \boldsymbol{y}) = \vec{\phi}(\boldsymbol{x}) \cdot \vec{\phi}(\boldsymbol{y})$ is equal to $(\boldsymbol{x} \cdot \boldsymbol{y})^2$. *(4 marks)*

$\overline{4}$

(c) Briefly describe the *random forest* algorithm. Explain why it is often very successful. *(5 marks)*

$\overline{5}$

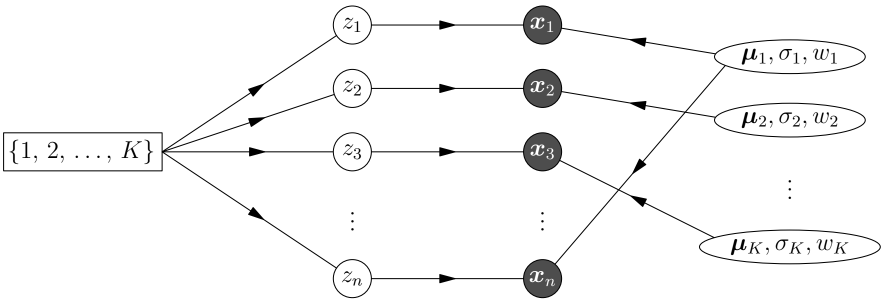
(d) Describe the difficulty of training a many layer multi-layer perceptron. *(5 marks)*

5

(e) Consider a mixture of Gaussians model for data $\mathcal{D} = \{\boldsymbol{x}_1,\, \boldsymbol{x}_2,\, \ldots,\, \boldsymbol{x}_n\}$. The model has parameters $\boldsymbol{\theta} = ((\boldsymbol{\mu}_1, \sigma_1, w_1),\, (\boldsymbol{\mu}_2, \sigma_2, w_2),\, \ldots,\, (\boldsymbol{\mu}_K, \sigma_K, w_K))$, such that the probability density for the data given the latent variables is

$$f(\mathcal{D}|\{z_1,\, z_2,\, \ldots,\, z_n\}) = \prod_{i=1}^{n} w_{z_i}\, \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_{z_i}, \sigma_{z_i}\, \mathbf{I}).$$

This can be represented by a graphical model



Draw the equivalent diagram using the plate notation. *(5 marks)*

5

(f) Show that the beta distribution $\mathrm{Beta}(p|a,b) = p^{a-1}(1-p)^{b-1}/B(a,b)$ is a conjugate prior to the binomial likelihood $\mathrm{Binom}(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k}$. Derive update equations for the parameters of the posterior distribution after observing k successes and $n-k$ failures.. *(5 marks)*

$\overline{5}$

End of question A1

Q1: (a) $\overline{6}$ (b) $\overline{4}$ (c) $\overline{5}$ (d) $\overline{5}$ (e) $\overline{5}$ (f) $\overline{5}$ Total $\overline{30}$

Section B

Question B 1

(a) Explain why choosing the maximum margin dividing plane is so important to the success of SVMs. *(5 marks)*

$\overline{}$ 5

(b) Sketch how slack variables, ξ_k, are introduced to allow some data points to lie within the margins. *(5 marks)*

$\overline{}$ 5

(c) Show

(i) how the constraints $y_k\left(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_k - b\right) \geq 1$ are changed by introducing the slack variables

(ii) how to modify the cost function $\frac{1}{2}\left\|\boldsymbol{w}\right\|^2$

(iii) the constraints on the slack variables.

Describe all the terms used. *(5 marks)*

i

ii

iii

$\overline{5}$

(d) Show how the Lagrangian, $\mathcal{L}$ is modified to include the slack variables and give the constraints on any Lagrange multipliers *(5 marks)*

$\overline{5}$

(e) By minimising with respect to the slack variables (i.e. setting $\frac{\partial \mathcal{L}}{\partial \xi_i} = 0$) obtain new constraints for the Lagrange multipliers α_i *(5 marks)*

$\overline{5}$

(f) Write down the general from for (i) a polynomial kernel and (ii) the radial basis function kernel *(5 marks)*

i

ii

$\overline{5}$

(g) Explain why it is important that a kernel is positive semi-definite and give three properties that a positive semi-definite kernel should have. *(5 marks)*

i

ii

iii

iv

$\overline{5}$

End of question B1

Q1: (a) $\frac{}{5}$ (b) $\frac{}{5}$ (c) $\frac{}{5}$ (d) $\frac{}{5}$ (e) $\frac{}{5}$ (f) $\frac{}{5}$ (g) $\frac{}{5}$ Total $\frac{}{35}$

Question B 2

(a) Sketch a typical CNN from taking in inputs to making a classification decision. Label the layers used. *(5 marks)*

5

(b) Briefly explain the following terms i) filters ii) feature maps iii) weight sharing iv) max pooling v) fully connected layer *(5 marks)*

i

ii

iii

iv

v

5

(c) Explain what is meant by i) Stochastic Gradient Descent ii) momentum in the context of learning and iii) mini-batches. *(5 marks)*

i

ii

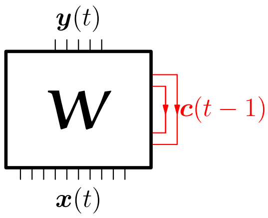
iii

5

(d) Briefly describe the motivation behind the design of Long-Short Term Memory (LSTM) units and how they achieve this. *(5 marks)*

5

(e) Consider a recurrent neural network with memory states $\boldsymbol{c}(t)$ as shown below.



Sketch how we can unroll the network in time to learn a sequence

$$(\boldsymbol{x}(1), \boldsymbol{y}(1)),\ (\boldsymbol{x}(2), \boldsymbol{y}(2)),\ \ldots,\ (\boldsymbol{x}(4), \boldsymbol{y}(4)).$$

(5 marks)

$\overline{}$ 5

(f) Explain what linear embedding units do and why they are so important in performing machine learning on languages. *(5 marks)*

$\overline{5}$

(g) Briefly explain the typical preprocessing steps that are carried out on documents before the data is feed into a learning machine. *(5 marks)*

$\overline{5}$

End of question B2

Q2: (a) $\frac{}{5}$ (b) $\frac{}{5}$ (c) $\frac{}{5}$ (d) $\frac{}{5}$ (e) $\frac{}{5}$ (f) $\frac{}{5}$ (g) $\frac{}{5}$ Total $\frac{}{35}$

Question B 3

(a) Explain for Gaussian Processes (GP) what is the prior, the likelihood and the posterior. *(5 marks)*

$\overline{5}$

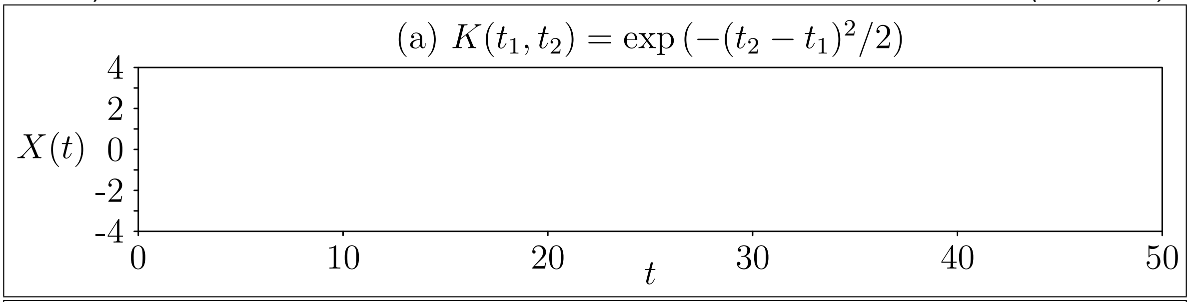
(b) Explain what the kernel function represents and how it could be measured empirically from many observations. *(5 marks)*

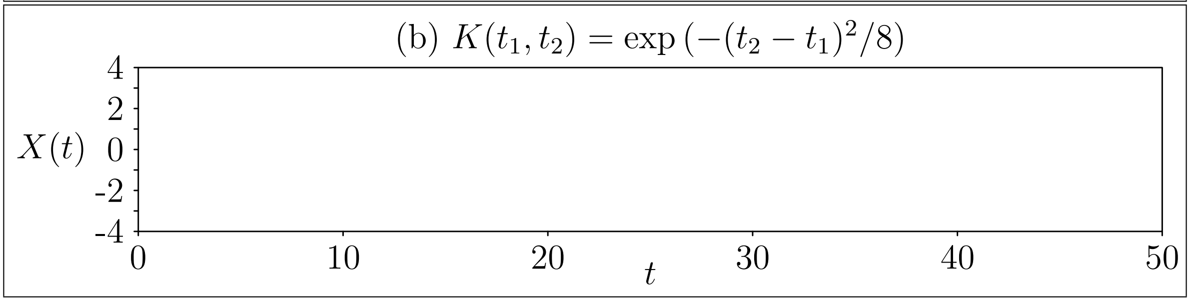
$\overline{5}$

(c) Consider a 1-d Gaussian Process, $X(t)$, with a kernel of the form

$$K(t_1, t_2) = \exp\left(-\frac{(t_2 - t_1)^2}{2\ell}\right).$$

Sketch three Gaussian Processes drawn from the prior with (a) $\ell = 1$ and (b) $\ell = 2$ (we are not looking for accuracy, but rather the effect of changing ℓ). *(5 marks)*





$\overline{5}$

(d) Explain the advantages and disadvantage of using the MAP solution rather than a full Bayesian solution. *(5 marks)*

5

(e) Explain why Monte Carlo techniques are often used to solve Bayesian inference problems. *(5 marks)*

$\overline{5}$

(f) Briefly describe in words the use of the MCMC algorithm in Bayesian inference. *(5 marks)*

$\overline{5}$

(g) When are probabilistic methods likely to give good results and what is the hurdle in using it? *(5 marks)*

$\overline{5}$

End of question B3

Q3: (a) $\frac{}{5}$ (b) $\frac{}{5}$ (c) $\frac{}{5}$ (d) $\frac{}{5}$ (e) $\frac{}{5}$ (f) $\frac{}{5}$ (g) $\frac{}{5}$ Total $\frac{}{35}$

END OF PAPER