UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2016/2017

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.

 Number of pages 16

Section A

Question A 1

(a) Explain what are (1) the **bias** and (2) the **variance** terms in the expected generalisation error and explain (3) the bias-variance dilemma. *(6 marks)*

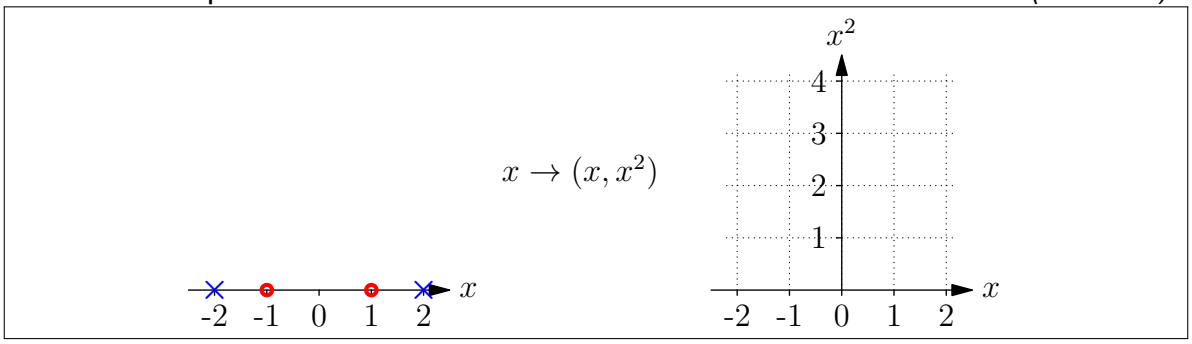
1

2

3

6

(b) For the one dimensional data points (crosses and circles) shown below, plot their position in an extended feature space created by the mapping $x \rightarrow (x, x^2)$. Draw the maximum margin dividing hyperplane in the extended feature space *(4 marks)*



4

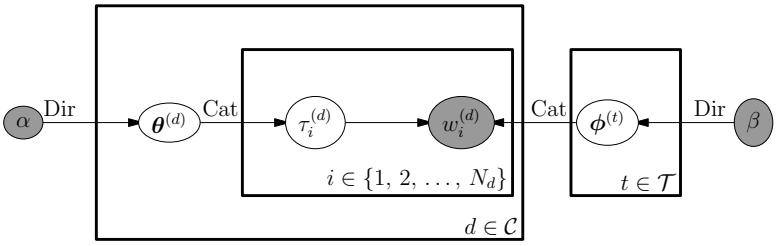
(c) Briefly describe the *Bagging* (bootstrap aggregating) algorithm, describe why it works, and give an example of a machine learning algorithm that uses it. *(5 marks)*

5

(d) Explain the difference between a discriminative probabilistic model and a generative model. Describe the advantages of each. *(5 marks)*

5

(e) The smoothed latent Dirichlet allocation topic model can be represented as a graphical model by the following plate diagram



where $\mathcal{C}$ is a set of documents and $\mathcal{T}$ is the set of topics. Sketch how documents of size N_d are generated by expanding the plate diagram to show the full word generation process. *(5 marks)*

5

(f) Show that the gamma distribution $\mathrm{Gam}(\mu|a,b) = b^a\,\mu^{a-1}\,\mathrm{e}^{-b\,\mu}/\Gamma(a)$ is a conjugate prior to the Poisson likelihood $\mathrm{Poi}(N|\mu) = \mu^N\,\mathrm{e}^{-\mu}/N!$ and derive the update equation for the parameters of the gamma distribution after observing N successes. *(5 marks)*

5

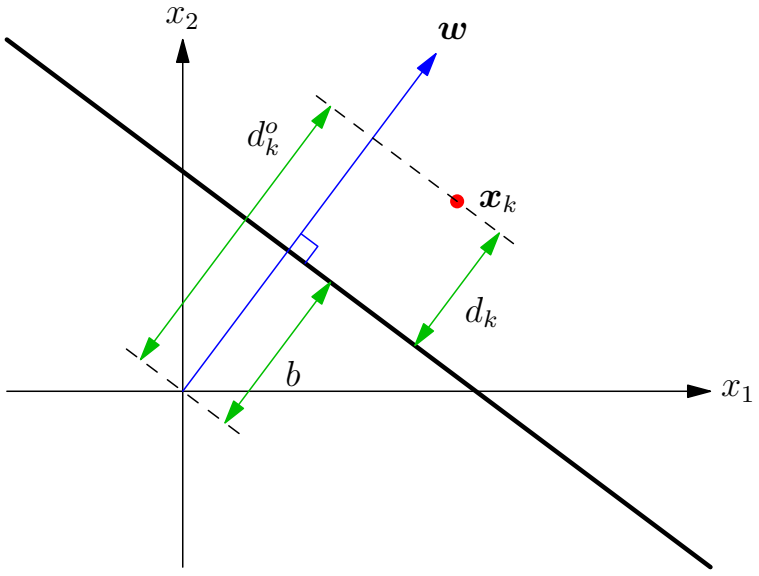
End of question A1

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

Section B

Question B 1

(a) Write down a formula for the minimum distance, d_k^0, between $\boldsymbol{x}_k$ and a hyperplane through the origin perpendicular to $\boldsymbol{w}$, and the minimum distance d_k from $\boldsymbol{x}_k$ to the hyperplane perpendicular to $\boldsymbol{w}$ displaced by b. *(5 marks)*



$d_k^o =$

$d_k =$

5

(b) Depending on the category $y_k \in \{-1, 1\}$, write down the condition for a data point to be at least a distance m above (or below if $y_k = -1$) the hyperplane shown in part (a). *(5 marks)*

5

(c) Define $\boldsymbol{w'} = \boldsymbol{w}/(m\|\boldsymbol{w}\|)$ and $b' = b/m$ to rewrite the condition from part (b) and explain why minimising $\|\boldsymbol{w}'\|^2$ is equivalent to maximising the margin m. *(5 marks)*

5

(d) Write down a Lagrangian for finding the maximal margin hyperplane for an SVM given data $(\boldsymbol{x}_k, y_k)$ for $k = 1, 2, \ldots, P$. *(5 marks)*

5

(e) Write down (1) the optimisation condition for the Lagrangian (i.e. what are you maximising or minimising with respect to) and (2) the conditions on the Lagrange multipliers. *(5 marks)*

1

2

5

(f) Find the weight vector $\boldsymbol{w}'$ and threshold b' which minimises the Lagrangian and by substituting the result back into the Lagrangian find the dual form for an optimisation problem. *(10 marks)*

$\overline{10}$

End of question B1

Q1: (a) $\overline{5}$ (b) $\overline{5}$ (c) $\overline{5}$ (d) $\overline{5}$ (e) $\overline{5}$ (f) $\overline{10}$ Total $\overline{35}$

• Do not write in this space •

Question B 2

(a) Explain through an example what idea computational learning theory is trying to capture *(5 marks)*

5

(b) Consider a finite set of hypotheses $\mathcal{H}$ for a binary classification task. If $h \in \mathcal{H}$ has an error rate ϵ, calculate the probability that it will make no error on P randomly selected (i.e. independent) patterns. *(5 marks)*

5

(c) Explain why the probability of any hypothesis with an error rate greater than ϵ will correctly classify P patterns is bounded by $|\mathcal{H}|\ \mathrm{e}^{-\epsilon P}$. *(5 marks)*

5

(d) Obtain a bound on the number of patterns required to ensure that a consistent learner (i.e. a machine that finds a hypothesis which is consistent with all the input patterns) will have an error less than ϵ with a probability of, at least, δ. *(5 marks)*

5

(e) Given a hypothesis space of with $|\mathcal{H}| = 10^{10}$ hypotheses, how many patterns do you need to guarantee an error rate less that 0.1% with a probability of at least 99.99%? *(5 marks)*

5

(f) Explain what the VC-dimension is and why it is needed. *(5 marks)*

5

(g) Explain why these bounds are of little value for understanding generalisation in deep learning. *(5 marks)*

5

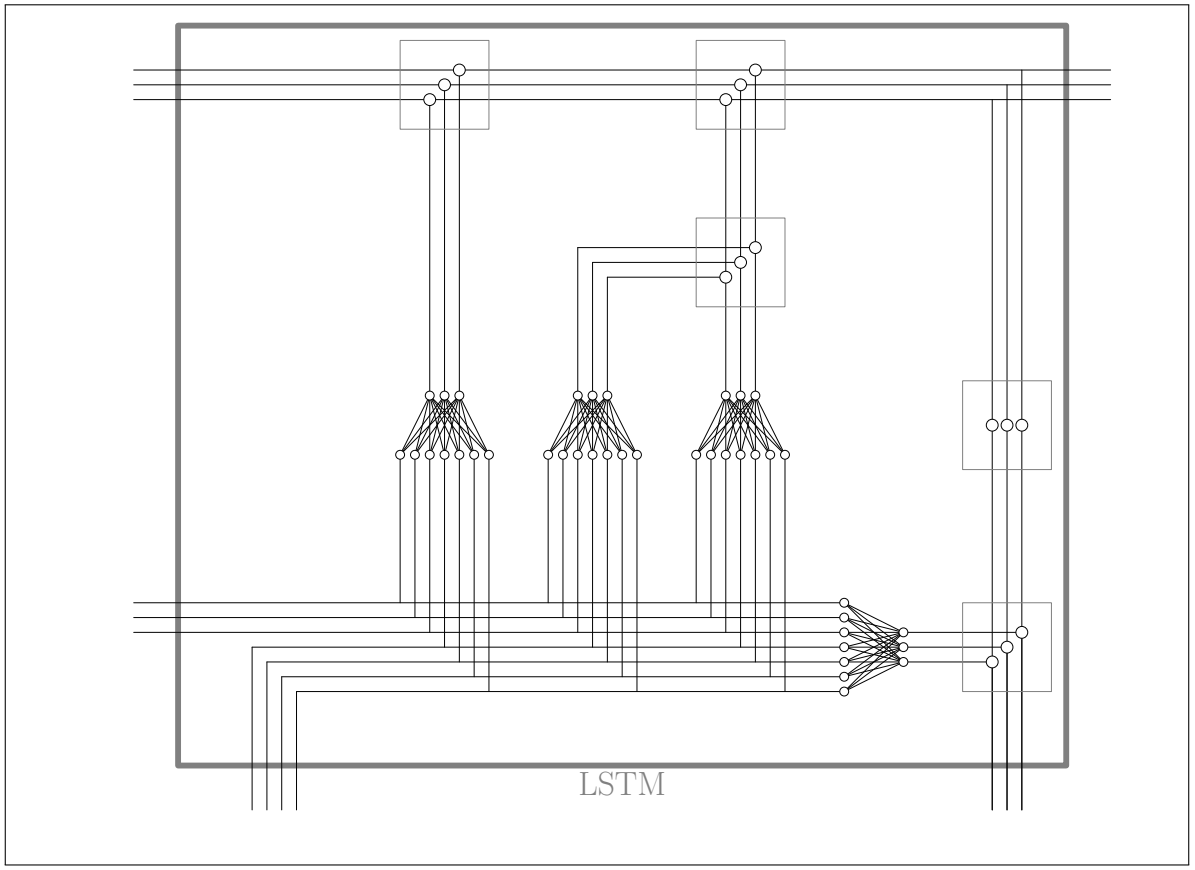
End of question B2

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

• Do not write in this space •

Question B 3

(a) Add annotations to the figure below of an LSTM showing i) the memory $c(t-1)$ and $c(t)$, ii) the input $x(t)$, iii) the output $y(t-1)$ and $y(t)$, iv) the forget gate, v) the input/update gate vi) the output gate. In addition show whether the gates are multiplicative or additive and whether the nodes are sigmoidal (σ) or tanh function. *(15 marks)*

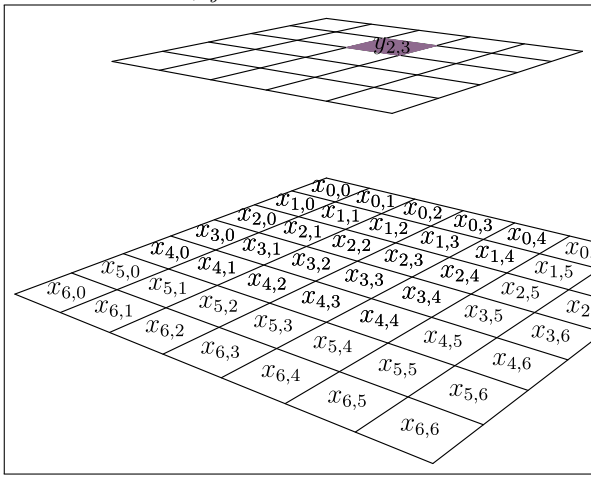


15

(b) Explain what problem LSTM were designed to solve and how their architecture solves these problems. *(5 marks)*

15

(c) In the figure shown below the bottom layer describes an image and the top a convolution layer. Show the pixels that would contribute to the 3×3 convolution at $y_{2,3}$. Write down the value of $y_{2,3}$ in terms of the convolution filter f_{δ_x,δ_y} and the image pixel values $x_{i,j}$. *(5 marks)*



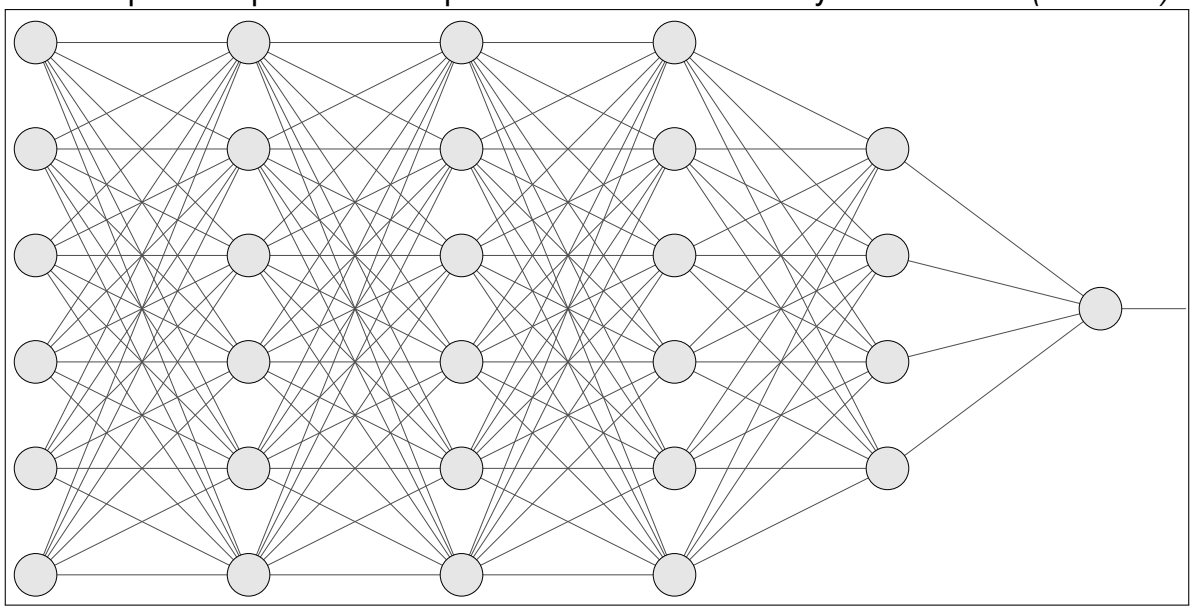
$y_{2,3} =$

15

(d) Sketch the architecture of a residual network and explain what this architecture allows. Why are they seen to work where traditional CNNs fail? *(5 marks)*

5

(e) In the diagram below of a deep network sketch a particular instance of dropout. Explain how dropout is carried out and why it is done? *(5 marks)*



5

End of question B3

Q3: (a) $\frac{}{15}$ (b) $\frac{}{5}$ (c) $\frac{}{5}$ (d) $\frac{}{5}$ (e) $\frac{}{5}$ Total $\frac{}{35}$

END OF PAPER