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)*

(Test of basic book knowledge.)

(i) The bias is the expected generalisation error averaged over machines trained on all possible data sets of a given size

(ii) The variance measures the expected variation from the average machine due to the fluctuations caused by using a finite training set

(iii) The bias-variance dilemma is that a simple machine is likely to have a high bias but low variance while a complex machine will have a low bias but high variance

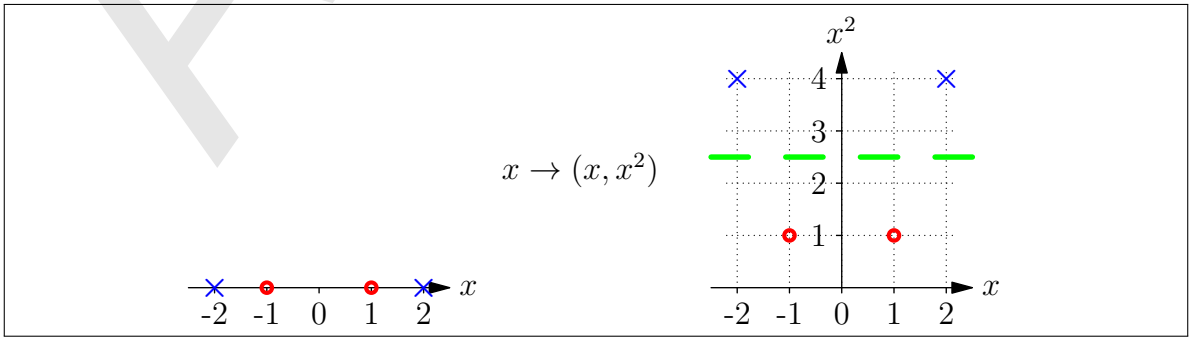
---

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)*

(This is an easy example of mapping data to an extended feature space, but one they have not seen before. It thus tests that they really understand.)

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)*

(Test basic knowledge)

Bagging is an ensemble learning technique that averages over a number of different machines. Each machine is trained on a different data set. The datasets are created by bootstrapping. That is, we sample the training set with replacement to create new training sets. We train a learning machine on each data set and then take the mean response of the set of machines.

This reduces the variance (in the bias-variance dilemma) through averaging over different datasets, thereby improving generalisation performance.

A prominent machine learning algorithm that uses bagging is the random forest algorithm that averages decision trees.

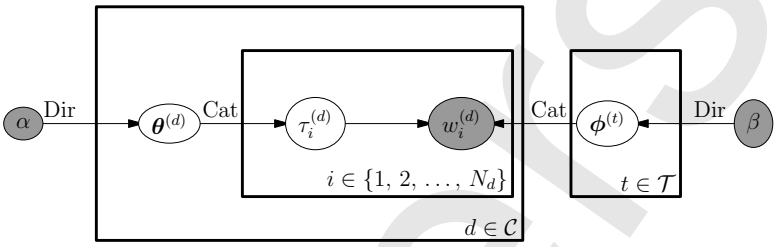
5

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

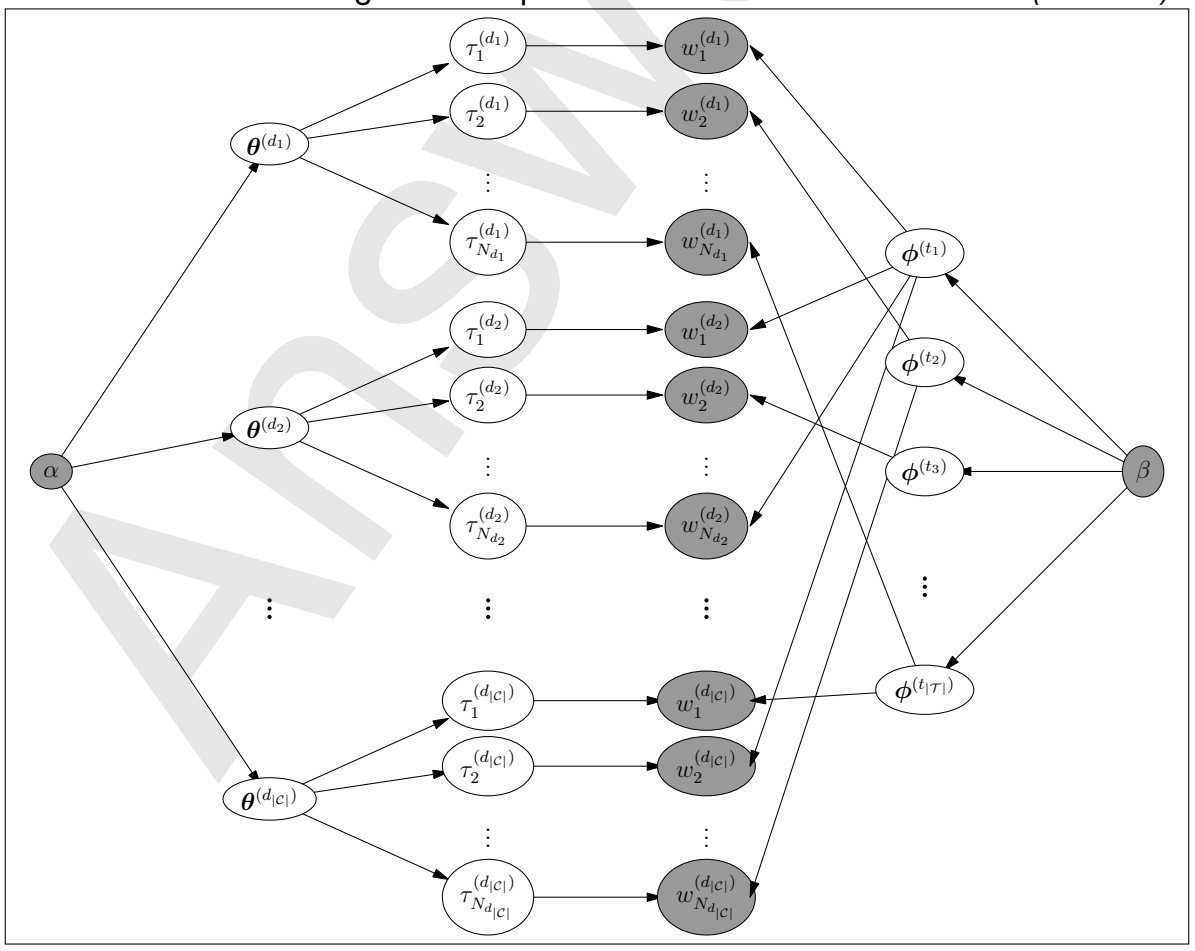
A discriminative probabilistic model predicts the probability of a target y given features $\boldsymbol{x}$, i.e. $\mathbb{P}(y|\boldsymbol{x})$. In generative models we generate both the target and the features using the joint probability $\mathbb{P}(y, \boldsymbol{x})$. Discriminative models are usually easier as we don't need to model the process of generating features. Generative models can provide a more accurate model and can be used in many different ways.

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)*

(Easy if you know what you are doing, but tests real understanding.)

We only need to consider the functional form with respect to μ. Thus the posterior is proportional to

$$f(\mu|N) \propto \mu^N\,\mathrm{e}^{-\mu}\mu^{a-1}\,\mathrm{e}^{-b\,\mu} \propto \mathrm{Gam}(\mu|a+N, b+1)$$

The updated equation is thus $(a,b) \rightarrow (a+N, b+1)$.

End of question A1

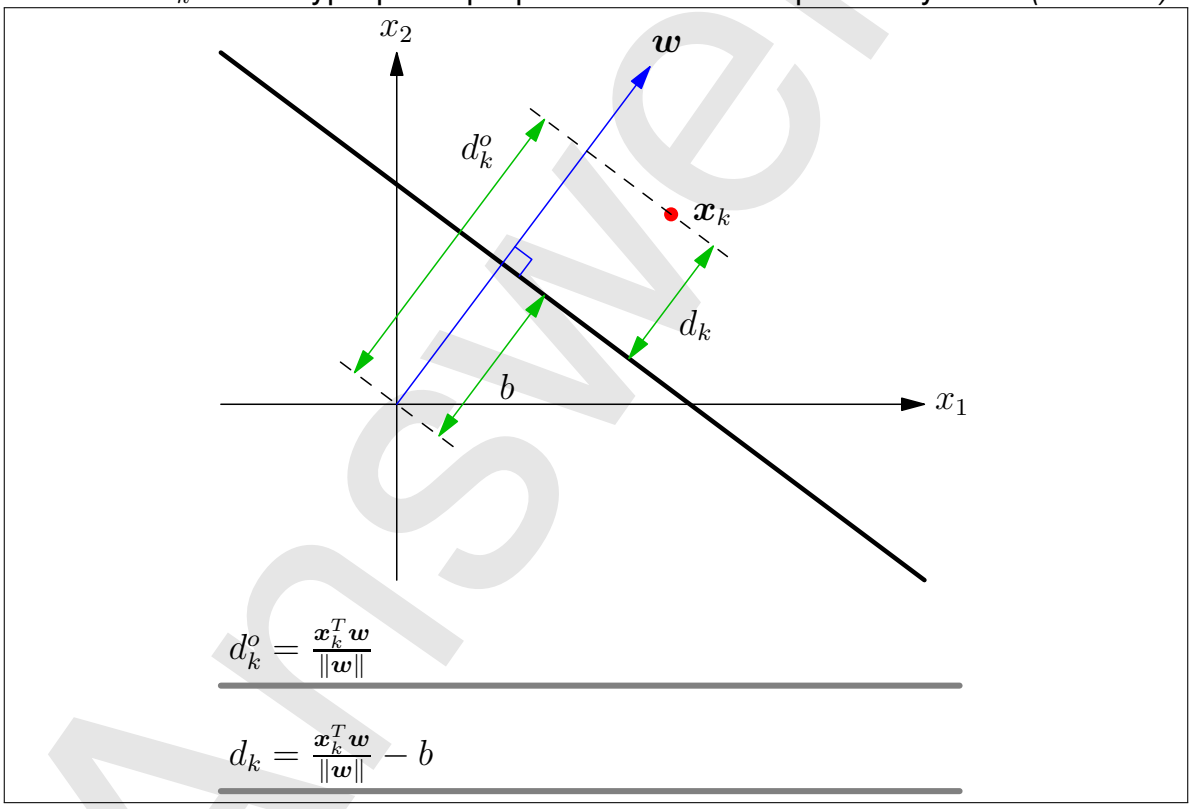
5

Q1: (a) ___/6	(b) ___/4	(c) ___/5	(d) ___/5	(e) ___/5	(f) ___/5	Total ___/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)*



$d_k^o = \frac{\boldsymbol{x}_k^T \boldsymbol{w}}{\|\boldsymbol{w}\|}$

$d_k = \frac{\boldsymbol{x}_k^T \boldsymbol{w}}{\|\boldsymbol{w}\|} - b$

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)*

$$y_k \left(\frac{\boldsymbol{x}_k^T \boldsymbol{w}}{\|\boldsymbol{w}\|} - b \right) \geq m$$

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)*

$$y_k \left(\boldsymbol{x}_k^T \boldsymbol{w}' - b'\right) \geq 1$$

5

$\|\boldsymbol{w}'\|^2 = 1/m^2$**, thus minimising** $\|\boldsymbol{w}'\|$ **is equivalent to maximising** $\boldsymbol{w'}$

(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)*

$$\mathcal{L}(\boldsymbol{w}', b', \boldsymbol{\alpha}) = \frac{\|\boldsymbol{w}'\|^2}{2} - \sum_{k=1}^{P} \alpha_k \left(y_k \left(\boldsymbol{x}_k^T \boldsymbol{w}' - b'\right) - 1\right)$$

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)*

(i) Optimisation condition

$$\min_{\boldsymbol{w}', b'} \max_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{w}', b', \boldsymbol{\alpha})$$

5

(ii) $\alpha_k \geq 0$ **for** $k = 1, 2, \ldots, P$

(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)*

Setting the derivatives with respect to w' to 0 we obtain

$$\nabla\mathcal{L} = \boldsymbol{w}' - \sum_{k=1}^{P} \alpha_k y_k \boldsymbol{x}_k = 0$$

or

$$\boldsymbol{w}' = \sum_{k=1}^{P} \alpha_k y_k \boldsymbol{x}_k$$

also

$$\frac{\partial\mathcal{L}}{\partial b} = \sum_{k=1}^{P} \alpha_k y_k = 0$$

substituting back into the Lagrangian

$$\mathcal{L} = -\frac{1}{2}\sum_{k,l=1}^{P} \alpha_k \alpha_l y_k y_l \boldsymbol{x}_k^T \boldsymbol{x_l} + \sum_{k,l=1}^{P} \alpha_k$$

The dual optimisation problem is

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\sum_{k,l=1}^{P} \alpha_k \alpha_l y_k y_l \boldsymbol{x}_k^T \boldsymbol{x_l} - \sum_{k,l=1}^{P} \alpha_k$$

$\overline{10}$

subject to

$$\alpha_k \geq 0 \quad \forall k = 1, \ldots, P \qquad\qquad \sum_{k=1}^{P} \alpha_k y_k = 0$$

End of question B1

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

• 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)*

(This is a conceptually challenging part of the course, although the mathematics is not hard. We start descriptively.)

In classification problems learning machines try to find a rule to separate training data into classes. However, there can be spurious rules for doing this. For example, in separating images of wolves and tigers we might find the background for the wolves are mostly white (or a particular pixel in the image is always white). By including more training examples we start to eliminate such spurious rules until the only remain rule is the correct one.

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)*

The probability of correctly classifying one pattern is $1 - \epsilon$. As the patterns are independent the probability of correctly classifying all of them is $(1 - \epsilon)^P$

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)*

The number of hypothesis with an error greater than ϵ is bounded above by the total number of hypotheses $|\mathcal{H}|$. The probability of any of them correctly classifying all the patterns is bounded by $(1-\epsilon)^P$. But, $1-\epsilon < \mathrm{e}^{-\epsilon}$ so the probability that any one of them correctly classifies all P patterns is strictly less than or equal to $|\mathcal{H}|\ \mathrm{e}^{-\epsilon P}$.

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)*

The probability of a consistent learner returning a hypothesis with an error greater than ϵ is $|\mathcal{H}|\ \mathrm{e}^{-\epsilon P}$. We require this probability to be less than δ. Taking logarithms (which does not change the inequality as it is monotonic), we find

$$\log\left(|\mathcal{H}|\right) - \epsilon\, P < \log\left(\delta\right)$$

or

$$P > \frac{1}{\epsilon}\left(\log\left(|\mathcal{H}|\right) + \log\left(\frac{1}{\delta}\right)\right).$$

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)*

(This is just substituting numbers into the formula, although it require understanding what these numbers mean.)

We have $\epsilon = 0.001$ and $\delta = 10^{-4}$, thus

$$P > 10^3 \left(\log\left(10^{10}\right) + \log\left(10^4\right)\right) = 14\,000\ \log(10) = 32\,236$$

5

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

The VC-dimension measures the capacity of a learning machine. That is the number of linearly separable patterns at which a learning machine cannot achieve all possible dichotomies of the inputs. This is used to obtain similar bounds to the one given above, but for machines with continuous parameters so that total number of possible hypotheses is infinite.

5

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

Many deep learning networks have been shown to perfectly learn a set of images with random labels (albeit that it takes longer than learning the correct labels). Thus the capacity of many deep learning machines is huge (VC-dimension effectively infinite). Computational learning theory thus provides no guarantees of generalisation. Since these guarantees are worst-case bound the fact that deep learning gives good generalisation does not contradict the theory.

End of question B2

5

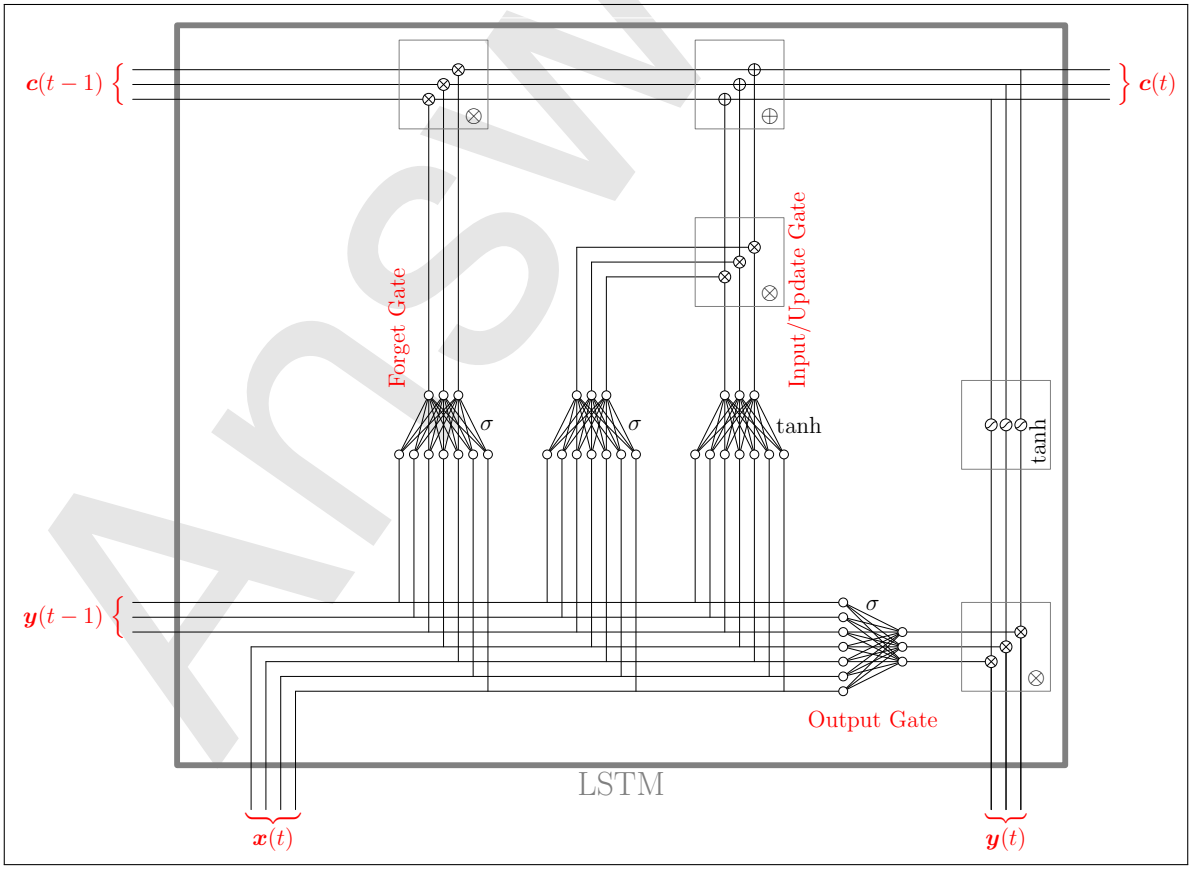
Q2: (a) $\frac{}{5}$ (b) $\frac{}{5}$ (c) $\frac{}{5}$ (d) $\frac{}{5}$ (e) $\frac{}{5}$ (f) $\frac{}{5}$ (g) $\frac{}{5}$ Total $\frac{}{35}$

• 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)*

(This is a non mathematical question, but tests a lot of knowledge about deep networks.)



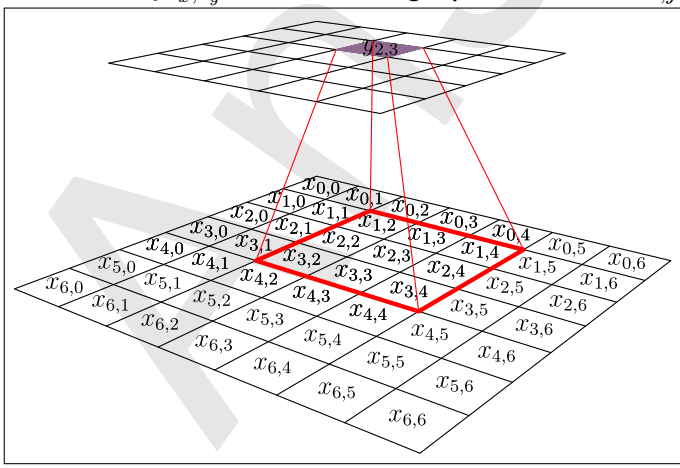
15

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

LSTM were designed to solve the vanishing/exploding gradient problem suffered by traditional recurrent networks. When dealing with very large series we may need to memorise events for many times steps. However, when we unwrap the training and backpropogate over many time steps the errors get multiplied together so with overwhelming probability will either vanishing or exploding (depending on the gain). The LSTM memory depends almost linear on the previous memory (up to a multiplicative factor which easily saturates at 1). This ensure that long term memories are relatively easy to learn.

5

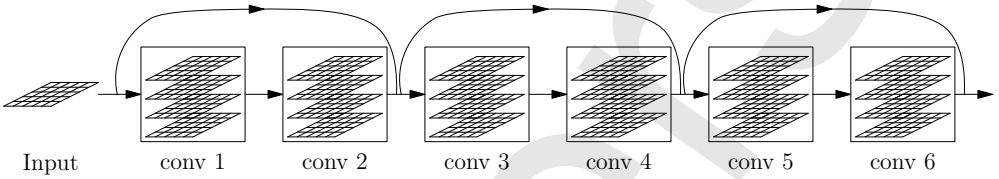
(c) In the figure shown below the bottom layer describes an image and the top a convolution layer. Show the pixels that would contribute to the 3×3 convolution at $y_{2,3}$. Write down the value of $y_{2,3}$ in terms of the convolution filter f_{δ_x,δ_y} and the image pixel values $x_{i,j}$. *(5 marks)*



$$y_{2,3} = f_{-1,-1}\,x_{1,2} + f_{-1,0}\,x_{1,3} + f_{-1,1}\,x_{1,4} + f_{0,-1}\,x_{2,2} + f_{0,0}\,x_{2,3} + f_{0,1}\,x_{2,4} + f_{1,-1}\,x_{3,2} + f_{1,0}\,x_{3,3} + f_{1,1}\,x_{3,4}$$

5

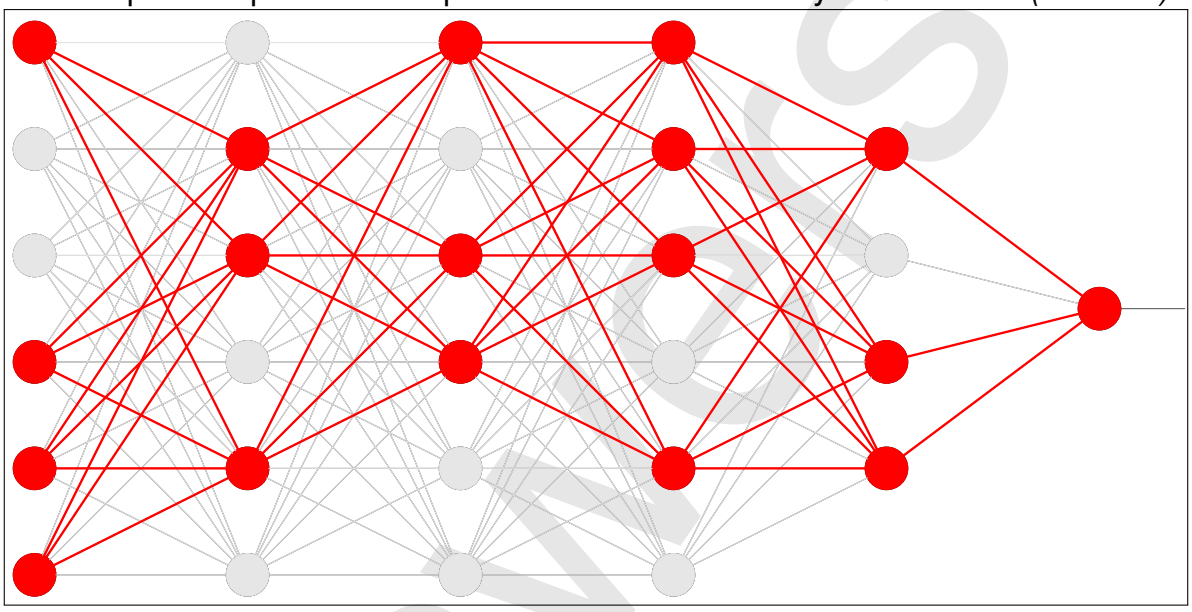
(d) Sketch the architecture of a residual network and explain what this architecture allows. Why are they seen to work where traditional CNNs fail? *(5 marks)*



Residual networks add skip connections between layers. They allow much deep networks to be trained. These tend to give better performance. The skip connections provide useful information in the deep part of the network helping the initial training. The connections also break the permutation and scaling symmetries between the filters which can substantially speed up learning.

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)*



Dropout selects some random nodes which it removes during a minibatch training run. At each minibatch (or even each training example) a new set of nodes is dropped. After training all nodes are restored with the weights reweighted. Dropout is a form of regularisation. It effectively, averages over (finds the geometric mean) the many sub-networks making up the full network, thus reducing the variance.

End of question B3

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

$\frac{}{5}$

END OF PAPER