UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2018 - 2019

ADVANCED MACHINE LEARNING

DURATION 120 MINS (2 Hours)

This paper contains 4 questions

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

An outline marking scheme is shown in brackets to the right of each question.

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

University approved calculators MAY be used.

A foreign language dictionary is permitted ONLY IF it is a paper version of a direct ‘Word to Word’ translation dictionary AND it contains no notes, additions or annotations.

18 page examination paper.

Section A

Question A1.

(a) Explain the meaning of (1) the *prior distribution*, $\mathbb{P}(x)$, for a random variable, X, and (2) the **likelihood**, $\mathbb{P}(\mathcal{D}|x)$, of the data, $\mathcal{D}$, and write down (3) *Bayes' rule* for the posterior.

Indicative Solution for Question A1(a).

(Test of basic book knowledge.)

1 **The prior encodes the prior belief about the distribution of the random variable before observing the data.**

2 **The likelihood is a model of the likelihood of the data given X.**

3 **Bayes' rule for the posterior, $\mathbb{P}(X|\mathcal{D})$, is given by**

$$\mathbb{P}(X|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|x)\ \mathbb{P}(x)}{\mathbb{P}(\mathcal{D})}$$

where

$$\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D}|x)\ \mathbb{P}(x)\ \mathrm{d}x$$

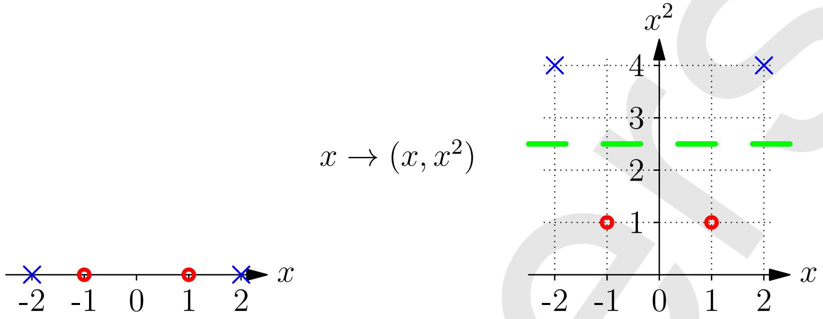
is a normalisation factor.

$\boxed{\overline{6}}$

[6 marks]

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

Indicative Solution for Question A1(b).



$\overline{4}$

[4 marks]

(c) Briefly describe the *Gradient Boosting* algorithm.

Indicative Solution for Question A1(c).

(Test basic knowledge)

Gradient boosting is a boosting algorithm where we build up a strong classifier, $\hat{f}(x)$, by adding weak classifies, $h_i(x)$,

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} h_i(\boldsymbol{x}).$$

In gradient boosting we greedily build a learning machine by adding a new weak learner trained on the residues in the error from the previous layer. To control over-fitting we often use a regulariser term and additionally we use earlier stopping based on the performance on a validation set. In many implementations such as XGBoost the weak learners of regression trees.

$\overline{5}$

[5 marks]

TURN OVER

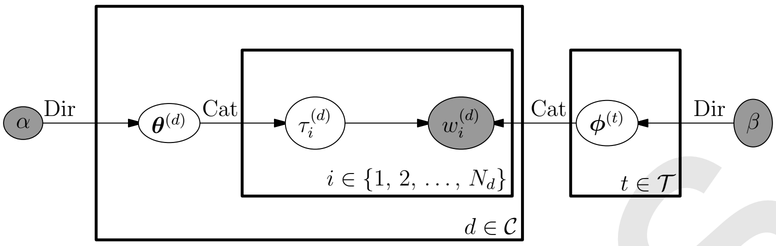
(d) Describe what you need to do practically to ensure that SVMs work well.

Indicative Solution for Question A1(d).

It is usually important to ensure that the features are normalised (e.g. by subtracting the mean and dividing by the standard deviation). It is also important to optimise over the punishment term, C, for the slack variables and over any parameters of the kernel (e.g. the γ variable in RBF kernels). Typically this involves performing a grid search over many orders of magnitude—often doubling the parameters when moving from one point of the grid to a neighbour. $\boxed{5}$

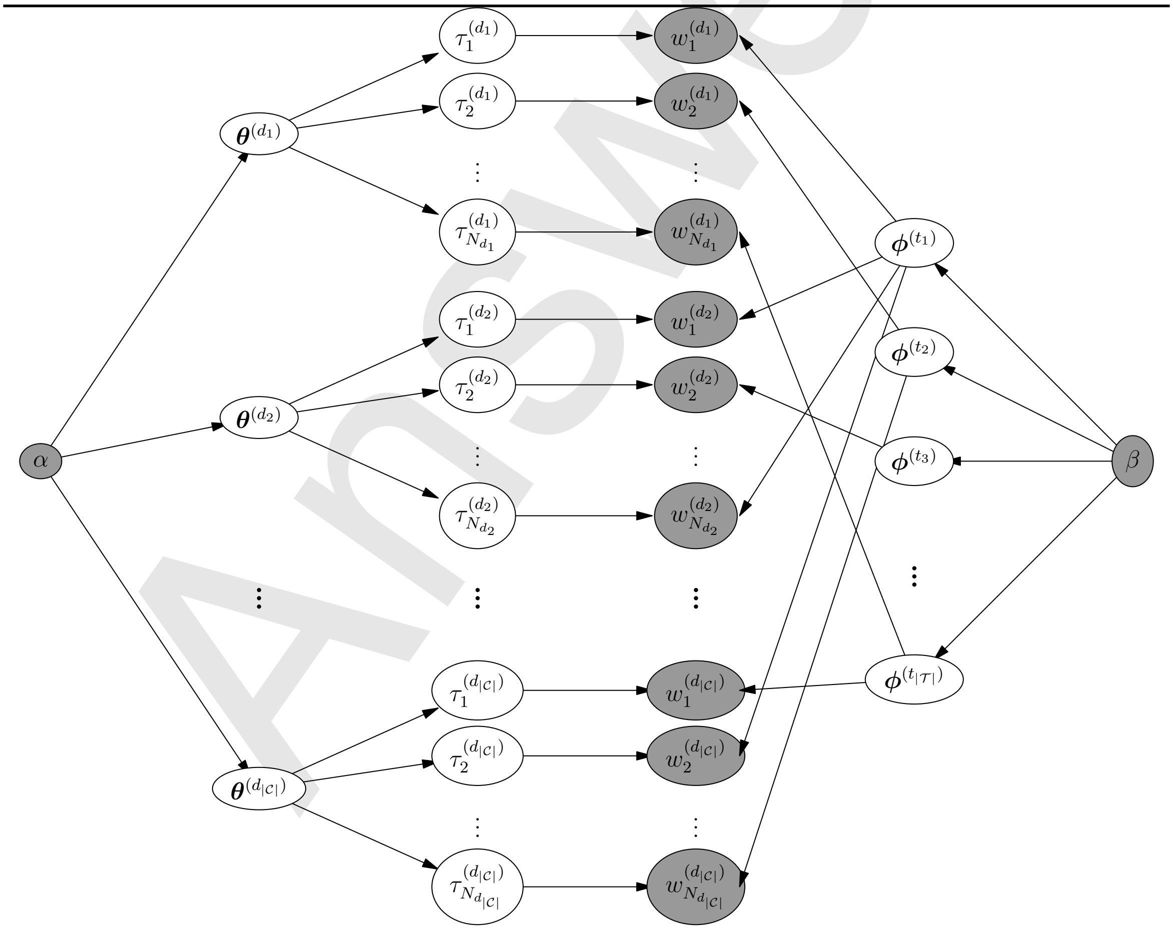
[5 marks]

(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.

Indicative Solution for Question A1(e).


α
$\boldsymbol{\theta}^{(d_1)}$
$\boldsymbol{\theta}^{(d_2)}$
$\boldsymbol{\theta}^{(d_{|\mathcal{C}|})}$
$\tau_1^{(d_1)}$
$\tau_2^{(d_1)}$
$\tau_{N_{d_1}}^{(d_1)}$
$\tau_1^{(d_2)}$
$\tau_2^{(d_2)}$
$\tau_{N_{d_2}}^{(d_2)}$
$\tau_1^{(d_{|\mathcal{C}|})}$
$\tau_2^{(d_{|\mathcal{C}|})}$
$\tau_{N_{d_{|\mathcal{C}|}}}^{(d_{|\mathcal{C}|})}$
$w_1^{(d_1)}$
$w_2^{(d_1)}$
$w_{N_{d_1}}^{(d_1)}$
$w_1^{(d_2)}$
$w_2^{(d_2)}$
$w_{N_{d_2}}^{(d_2)}$
$w_1^{(d_{|\mathcal{C}|})}$
$w_2^{(d_{|\mathcal{C}|})}$
$w_{N_{d_{|\mathcal{C}|}}}^{(d_{|\mathcal{C}|})}$
$\boldsymbol{\phi}^{(t_1)}$
$\boldsymbol{\phi}^{(t_2)}$
$\boldsymbol{\phi}^{(t_3)}$
$\boldsymbol{\phi}^{(t_{|\mathcal{T}|})}$
β

[5 marks]

5

TURN OVER

(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.

Indicative Solution for Question A1(f).

(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 $(a,b) \to (a+N, b+1)$.

$\boxed{\frac{\quad}{5}}$

[5 marks]

End of question A1

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

Section B

Question B1.

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

Indicative Solution for Question B1(a).

(Conceptually challenging. This question is from last year, but almost no one tackled it.)

The prior is a measure over function such that the probability of points in space are normally distributed with a two-point correlation function given by the kernel $K(\boldsymbol{x}, \boldsymbol{y})$—that is it is a Gaussian Process. The likelihood is typically a Gaussian between the observed points and the prediction of the Gaussian process. The posterior is also a Gaussian Process conditioned on the observations.

$\boxed{\bar{5}}$

[5 marks]

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

Indicative Solution for Question B1(b).

The kernel give the two point correlation function. If we have enough data point $(\boldsymbol{x}_i, y_i)$ we can compute the pairwise distances $d = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. We can then compute the mean correlations $(y_i - \mu)(y_i - \mu)/\sigma^2$ for all pairs of points whose distance lies within some narrow interval. Note here μ and σ^2 are the empirical means and variances of the complete set of target values y_i. This correlation as a function of the distance function should be similar to the kernel function. Furthermore, within each interval the data should be roughly normally distributed.

$\boxed{\bar{5}}$

[5 marks]

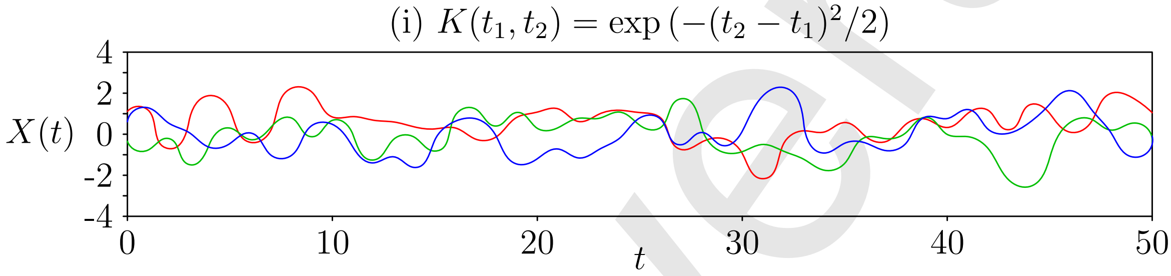
TURN OVER

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

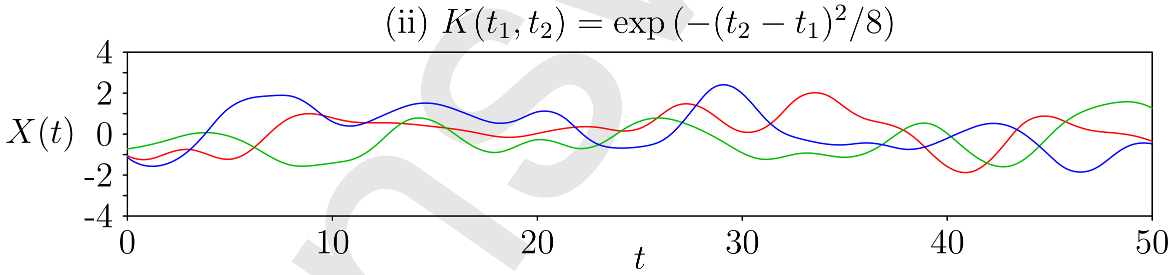
$$K(t_1, t_2) = \exp\left(-\frac{(t_2 - t_1)^2}{2\ell}\right).$$

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

Indicative Solution for Question B1(c).


(i) $K(t_1, t_2) = \exp\left(-(t_2 - t_1)^2/2\right)$

Indicative Solution for Question B1(c).


(ii) $K(t_1, t_2) = \exp\left(-(t_2 - t_1)^2/8\right)$

$\overline{5}$

[5 marks]

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

Indicative Solution for Question B1(d).

The MAP solution is usually much simpler to calculate, particularly when there is no closed form solution for the posterior. Also we don't need to evaluate the normalisation term which often can only be computed approximately with much effort. On the other hand, if the posterior is not unimodal and strongly concentrated around its maximum value then the MAP solution can be very inaccurate. Note also that the MAP solution doesn't give a probabilistic answer so we cannot, for example, compute the accuracy of the prediction (something you can do in the full Baysian framewwork).

$\boxed{\bar{5}}$

[5 marks]

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

Indicative Solution for Question B1(e).

For many Bayesian problems the posterior has no closed form solution so it cannot be expressed. Monte Carlo techniques allow us to draw samples from the posterior which allow us to compute many quantities of interest such as the posterior mean (often the best single prediction) and the posterior variance (giving an indication of the expected error).

$\boxed{\frac{\ }{5}}$

[5 marks]

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

Indicative Solution for Question B1(f).

In MCMC we explore the posterior distributions by making small jumps in the possible solutions with a probability that satisfies detail balance. That is, the probability of making a move is dependent on the ratio of the posterior probability of the solution before and after the move. This ensures that the distributions of sampled points converges to the posterior distribution. We need to throw away the initial points (burn-in phase). This gives us sample points from the posterior (although to be independent we need to wait for the points to decorrelate).

$\boxed{\frac{\ }{5}}$

[5 marks]

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

Indicative Solution for Question B1(g).

Probabilistic methods are optimal when we have an accurate model of the likelihood of the data and a good posterior. This is typically the case when we have a good understanding of how the data is generated. Probabilistic methods are expensive to use as we have to carefully model the likelihood and prior. Often we have little understanding of what generated the data, although we may have some expectation about the solution (e.g. it is likely to be continuous and not rapidly changing although these are often hard to specify as a prior). In such cases probabilistic methods are often dominated by other (more generic) machine learning techniques.

$\frac{\quad}{5}$

[5 marks]

End of question B1

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

Question B2.

(a) Show that the expected generalisation given by

$$\mathbb{E}_{\mathcal{D}}\big[E(\mathcal{D})\big] = \mathbb{E}_{\mathcal{D}}\left[\sum_{x \in \mathcal{X}} p(\boldsymbol{x}) \left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - f(\boldsymbol{x})\right)^2\right]$$

can be written as the sum of a bias term, B, and variance term V where

$$B = \sum_{x \in \mathcal{X}} p(\boldsymbol{x}) \left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2, \quad V = \mathbb{E}_{\mathcal{D}}\left[\sum_{x \in \mathcal{X}} p(\boldsymbol{x}) \left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right)^2\right]$$

where $\hat{f}_m(\boldsymbol{x}) = \mathbb{E}_{\mathcal{D}}\Big[\hat{f}(\boldsymbol{x}|\mathcal{D})\Big]$ is the prediction made by averaging over all machines.

Indicative Solution for Question B2(a).

(Students find this derivation difficult because it requires understanding expectations. Being rather conceptual students often struggle with this.)

The first step is to subtract and add the response of the mean machine

$$\begin{aligned}
\bar{E}_G = \mathbb{E}_{\mathcal{D}}\big[E_G(\mathcal{D})\big] &= \mathbb{E}_{\mathcal{D}}\left[\sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) \left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - f(\boldsymbol{x})\right)^2\right] \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x})\, \mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - f(\boldsymbol{x})\right)^2\right] \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x})\, \mathbb{E}_{\mathcal{D}}\left[\left(\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right) + \left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)\right)^2\right] \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) \Bigg(\mathbb{E}_{\mathcal{D}}\left[\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right)^2 + \left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)^2\right] \\
&\qquad\qquad + \mathbb{E}_{\mathcal{D}}\Big[2\left(\hat{f}(\boldsymbol{x}|\mathcal{D}) - \hat{f}_m(\boldsymbol{x})\right)\left(\hat{f}_m(\boldsymbol{x}) - f(\boldsymbol{x})\right)\Big]\Bigg)
\end{aligned}$$

The last term vanishes on taking the expectation. This leaves us with the variance term plus the bias.

$\boxed{\overline{15}}$

[15 marks]

(b) Explain the *bias* and the *variance* terms in words.

Indicative Solution for Question B2(b).

(2 marks each and an extra mark if both are correct.)

(i) The *bias* is the generalisation performance of the mean machine (i.e. the prediction made by taking the mean response averaged over machines trained with every possible training set of a given size).

(ii) The *variance* measures how the response of the individual machines vary from machine to machine.

$\overline{5}$

[5 marks]

(c) What is the bias-variance dilemma.

Indicative Solution for Question B2(c).

To get good generalisation we want to both reduce the bias and variance. To reduce the bias we need a complex machine, but that will typically to increase the variance. Conversely a simple machine is likely to have a small variance, but a high bias.

$\overline{5}$

[5 marks]

(d) Explain how the *random forest* algorithm attempts to overcome the bias-variance dilemma.

Indicative Solution for Question B2(d).

Random forest averages a large number of decision trees that have been trained on slightly different data sets (through bootstrapping) using a different set of features. This averaging reduces the variance.

$\boxed{\frac{}{5}}$

[5 marks]

(e) Explain how regularisation helps in the context of the bias-variance dilemma.

Indicative Solution for Question B2(e).

By introducing a regularisation term we can use powerful machines (with potentially low bias), however, the regulariser usually makes the learning machine less sensitive to the data, thus reducing the variance.

$\boxed{\frac{}{5}}$

[5 marks]

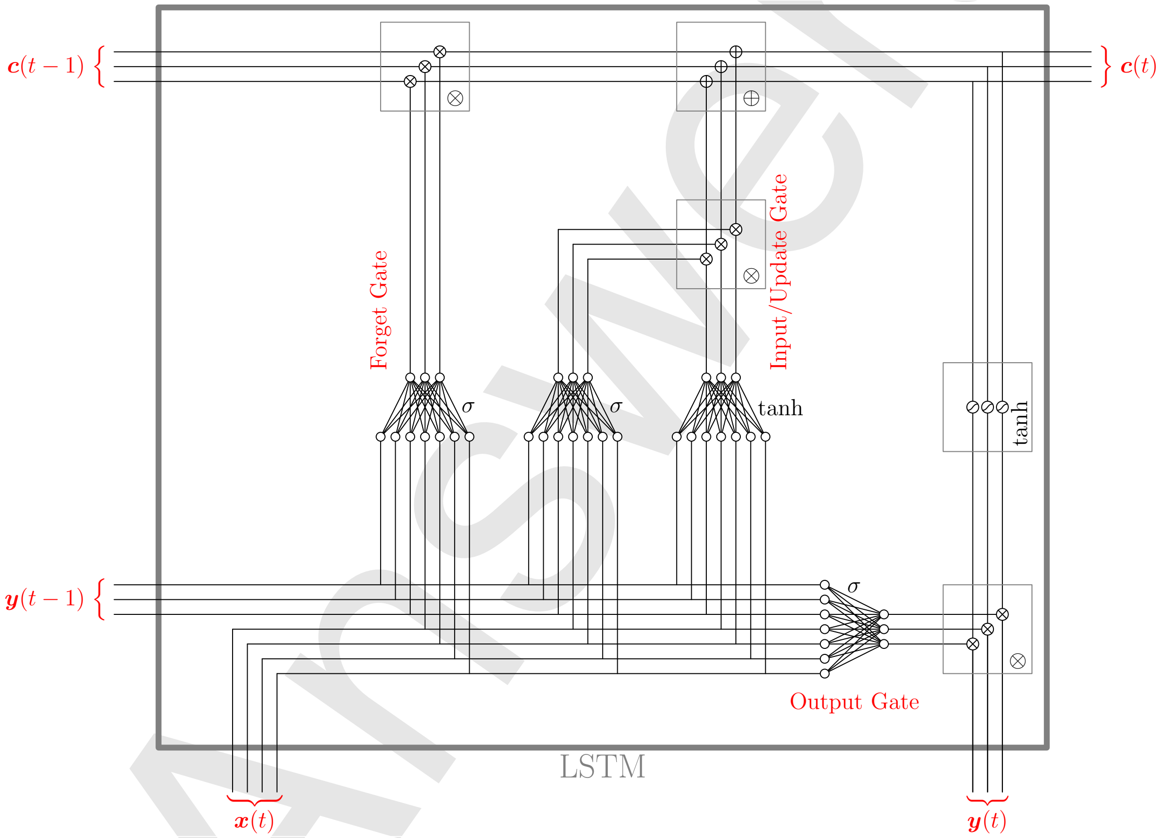
End of question B2

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

Question B3.

(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.

Indicative Solution for Question B3(a).



$\overline{15}$

[15 marks]

TURN OVER

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

Indicative Solution for Question B3(b).

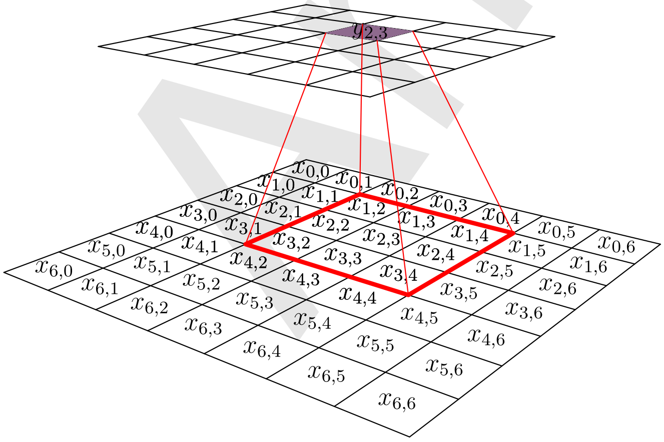
LSTM were designed to solve the vanishing/exploding gradient problem suffered by traditional recurrent networks. When dealing with very large series we may need to memorise events for many times steps. However, when we unwrap the training and backpropogate over many time steps the errors get multiplied 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.

$\overline{5}$

[5 marks]

(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 $x_{i,j}$.

Indicative Solution for Question B3(c).


$y_{2,3}$
$x_{0,0}$ $x_{0,1}$ $x_{0,2}$ $x_{0,3}$ $x_{0,4}$ $x_{0,5}$ $x_{0,6}$
$x_{1,0}$ $x_{1,1}$ $x_{1,2}$ $x_{1,3}$ $x_{1,4}$ $x_{1,5}$ $x_{1,6}$
$x_{2,0}$ $x_{2,1}$ $x_{2,2}$ $x_{2,3}$ $x_{2,4}$ $x_{2,5}$ $x_{2,6}$
$x_{3,0}$ $x_{3,1}$ $x_{3,2}$ $x_{3,3}$ $x_{3,4}$ $x_{3,5}$ $x_{3,6}$
$x_{4,0}$ $x_{4,1}$ $x_{4,2}$ $x_{4,3}$ $x_{4,4}$ $x_{4,5}$ $x_{4,6}$
$x_{5,0}$ $x_{5,1}$ $x_{5,2}$ $x_{5,3}$ $x_{5,4}$ $x_{5,5}$ $x_{5,6}$
$x_{6,0}$ $x_{6,1}$ $x_{6,2}$ $x_{6,3}$ $x_{6,4}$ $x_{6,5}$ $x_{6,6}$

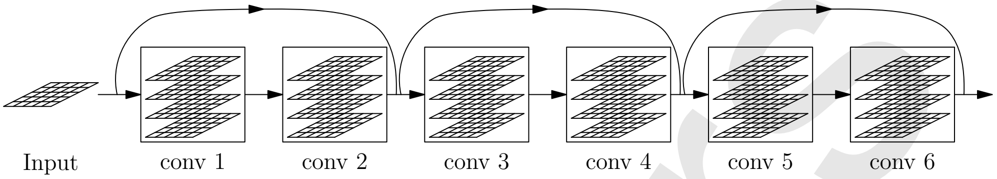
$$\begin{aligned} y_{2,3} = \; & f_{-1,-1}\, x_{1,2} + f_{-1,0}\, x_{1,3} + f_{-1,1}\, x_{1,4} \\ & + f_{0,-1}\, x_{2,2} + f_{0,0}\, x_{2,3} + f_{0,1}\, x_{2,4} \\ & + f_{1,-1}\, x_{3,2} + f_{1,0}\, x_{3,3} + f_{1,1}\, x_{3,4} \end{aligned}$$

$\overline{5}$

[5 marks]

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

Indicative Solution for Question B3(d).



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.

$\boxed{\bar{5}}$

[5 marks]

TURN OVER

(e) Describe what is meant by transfer learning in the context of CNNs.

Indicative Solution for Question B3(e).

Transfer learning is when we use the weights of a CNN trained on one dataset (often ImageNet) on another task. In this case we usually replace the layers at the end of the network (often the fully connected layers), but keep the majority of the CNN filters. Typically, we attach a new head to the network and retrain the head on a new (and often much smaller) data set.

$\frac{\ }{5}$

[5 marks]

End of question B3

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

END OF PAPER