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.

 Number of pages 19

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)*

(Many different answers to this, e.g.)

(i) SVMs excel with small hig-dimensional data sets (well balanced, no missing data)

(ii) Gradient boosting is good with messy tabular data (missing data, mixed types, etc.)

(iii) CNNs excel with images and signals with very large training sets.

6

(b) Show that for the mapping

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

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)*

(This is an extension of the 2-D example shown in the lecturer.)

$$\begin{aligned} K(\boldsymbol{x}, \boldsymbol{y}) &= \vec{\phi}(\boldsymbol{x}) \cdot \vec{\phi}(\boldsymbol{y}) \\ &= (x_1^2, x_2^2, x_3^2, \sqrt{2}\, x_1\, x_2,\ \sqrt{2}\, x_1\, x_3,\ \sqrt{2}\, x_2\, x_3) \\ &\quad \cdot (y_1^2, y_2^2, y_3^2, \sqrt{2}\, y_1\, y_2,\ \sqrt{2}\, y_1\, y_3,\ \sqrt{2}\, y_2\, y_3) \\ &= x_1^2 y_1^2 + x_2^2 y_2^2 + x_3^2 y_2^2 + 2\, x_1\, y_1\, x_2\, y_2 + 2\, x_1\, y_1\, x_3\, y_3 + 2\, x_2\, y_2\, x_3\, y_3 \\ &= (x_1\, y_1 + x_2\, y_2 + x_3\, y_3)^2 = (\boldsymbol{x} \cdot \boldsymbol{y})^2 \end{aligned}$$

4

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

(Test basic knowledge) **Random forest uses an ensemble of decision trees. It uses bootstrap aggregation (bagging) to obtain different trees. However, it also selects a relatively small number of variables (typically between 1 and $\sqrt{p}$ where p is the number of features) and constructs shallow trees. This creates quite uncorrelated trees. By averaging a large number of such trees it reduces the variance. Although each machine is not particular good (weak learners) the average machine can approximate quite complex decision surfaces (i.e. have low bias).**

5

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

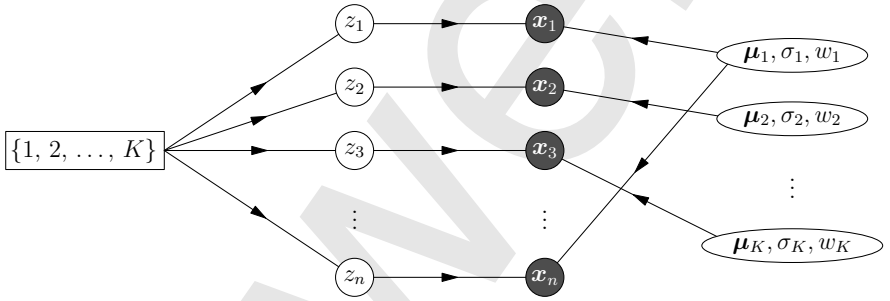
In MLPs with a large number of layers it is difficult to prevent both the forward responses and the back-propagated errors from either vanishing or exploding. Furthermore, there is a huge permutation symmetry between the nodes in each layer meaning that there are a huge number of ambiguous directions in weight space where there are optimal solution. It is typically very hard to train the early layers in a network with more than two hidden layers.

5

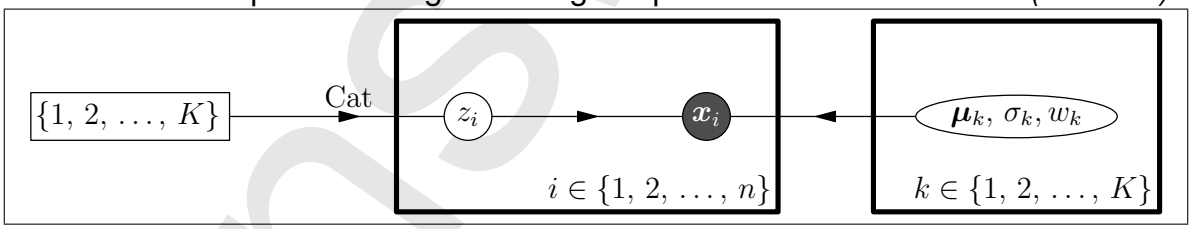
(e) Consider a mixture of Gaussians model for data $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. The model has parameters $\boldsymbol{\theta} = ((\boldsymbol{\mu}_1, \sigma_1, w_1), (\boldsymbol{\mu}_2, \sigma_2, w_2), \ldots, (\boldsymbol{\mu}_K, \sigma_K, w_K))$, 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)*

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

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

$$f(p|k) \propto p^k (1-p)^{n-k} \times p^{a-1}(1-p)^{b-1} \propto \mathrm{Beta}(p|a+k, b+n-k)$$

The updated equation is thus $(a,b) \rightarrow (a+k, b+n-k)$.

End of question A1

5

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

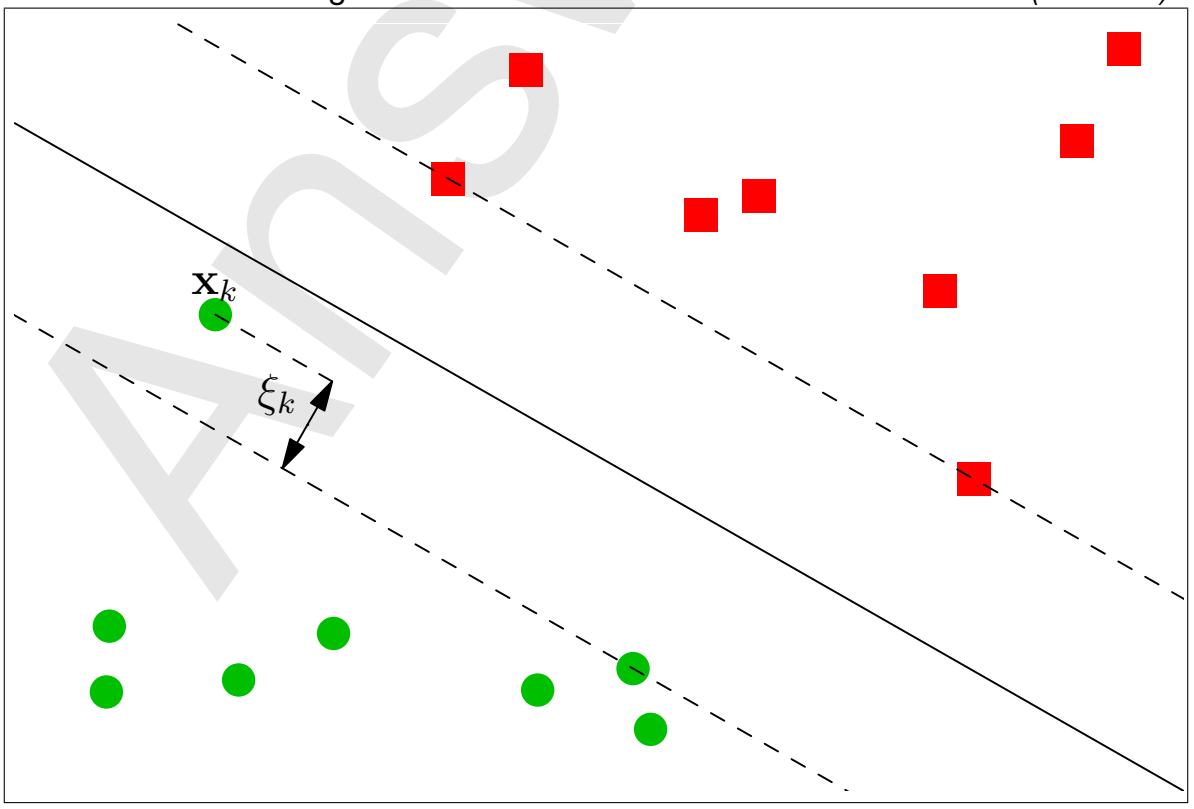
(a) Explain why choosing the maximum margin dividing plane is so important to the success of SVMs. *(5 marks)*

(Start easily, but important to understand.)

By choosing the maximum margin hyperplane the SVM finds a solution that tends to generalise well. That is, it is the dividing plane that can tolerate the largest level of noise on the data points and still give the correct classification. This is essential when using an extended feature space which would otherwise have a huge capacity to over-fit the data set.

5

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



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) $y_k\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_k - b\right) \geq 1 - \xi_k$

(ii)

$$\frac{1}{2}\left\|\boldsymbol{w}\right\|^2 + C\sum_{i=1}^{P}\xi_i$$

where C is a parameter that has to be chosen (usually by grid search) and P is the number of training patterns.

(iii) $\xi_i \geq 0 \quad \forall\, i \in \{1,\, 2,\, \ldots,\, P\}$

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)*

$$\mathcal{L} = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{k=1}^{n}\xi_k - \sum_{k=1}^{P}\alpha_k\left(y_k\,\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_k - b\right) - 1 + \xi_k\right) - \sum_{k=1}^{P}\beta_k\,\xi_k$$

subject to $\alpha_i, \beta_i \geq 0$**.**

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)*

$$\frac{\partial \mathcal{L}}{\partial \xi_k} = C - \alpha_k - \beta_k = 0$$

5

But $\beta_k \geq 0$ this implies $\alpha_k \leq C$ so that $0 \leq \alpha_k \leq C$ (since $\alpha_k \geq 0$ as part of the KKT conditions) for all $k = 1, 2, \ldots, n$.

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

(i) The general form of the polynomial kernel is $(\boldsymbol{x}^\mathsf{T}\boldsymbol{y})^d$ (or $(1 + \boldsymbol{x}^\mathsf{T}\boldsymbol{y})^d$)

(ii) The general form of the RBF kernel is $\exp(-\gamma\|\boldsymbol{x} - \boldsymbol{y}\|^2)$.

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) Kernel functions need to be positive semi-definite so that they have sensible (non-negative) distances. That is the margins are positive. (2 points)

(ii) The eigenvalues of a positive semi-definite kernel function are non-negative

(iii) A positive semi-definite kernel function can always be written as

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \phi_i(\boldsymbol{x})\, \phi_i(\boldsymbol{y})$$

for some set of real functions $\phi_i(\boldsymbol{x})$

(iv) The quadratic form satisfies

$$\int f(\boldsymbol{x})\, K(\boldsymbol{x}, \boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y} \geq 0$$

for any real function $f(\boldsymbol{x})$**.**

5

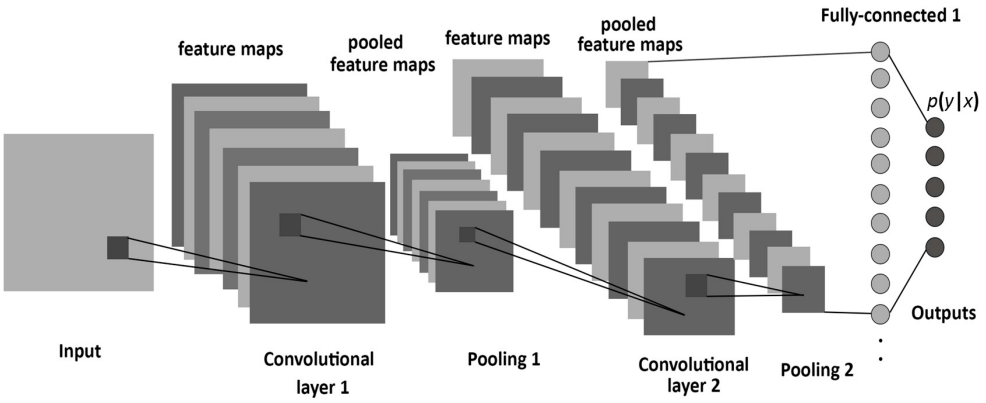
End of question B1

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)*

(Looking for some convolutional layers, pooling layers and fully connect layers and/or softmax layer.)



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) Filters are the weighted patches that are moved across the images or feature maps

(ii) feature maps are the result from the filters obtained at each location

(iii) weight sharing is the term used to indicate the same weights (filters) are used many times to produce a feature map

(iv) max pooling is a technique to reduce the size of the feature map by choosing the largest response in a small (usually 2×2) block

(v) fully connected layers are traditional MLP type layers where every neuron receives inputs from the same set of nodes

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) Stochastic gradient descent means that we compute the gradient and update the weights after each image (or mini-batch)—2 points

(ii) Momentum is when our update "velocity" remembers previous updates that are slowly changed by the current gradient—2 points

(iii) Mini-batches is when we compute the average gradient for a small number of input patterns (e.g. 5–100) at each time step and update according to this average gradient—1 point.

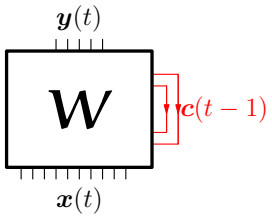
5

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

LSTMs are designed to make it easy for memories to be kept for many epochs. That is the amplification factor is 1. This is achieved by using multiplicative gates with sigmoid activation units that easily saturate so that the memory is maintained. This prevents the vanish (or exploding) problem when a signal is feedback over many epochs.

5

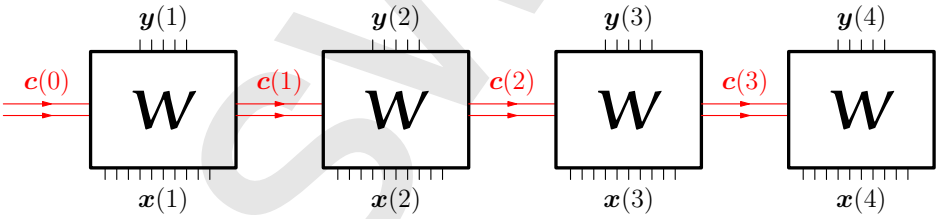
(e) Consider a recurrent neural network with memory states $\boldsymbol{c}(t)$ as shown below.



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)



5

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

Linear embeddings map from large dimensional one-hot encodings to a much lower dimensional space such that similar words are mapped to the same region in the embedding space. The relative positions often captures some semantic meaning in the words.

It is important in language models because languages usually have a very large vocabulary. A direct encoding would have a huge number of inputs which most learning machines struggle with.

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)*

This can involve an number of stages. E.g. splitting into words (tokenising), often removing of punctuation, setting to lower case, stemming, and removal of stop words. Sometimes, the document might be represented as a bag of words (using a code book) where often TF-IDF is used to score the words or word embeddings are used.

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 $\frac{}{35}$

Question B 3

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

(Conceptually challenging.)

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 $K(x, y)$—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.

5

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

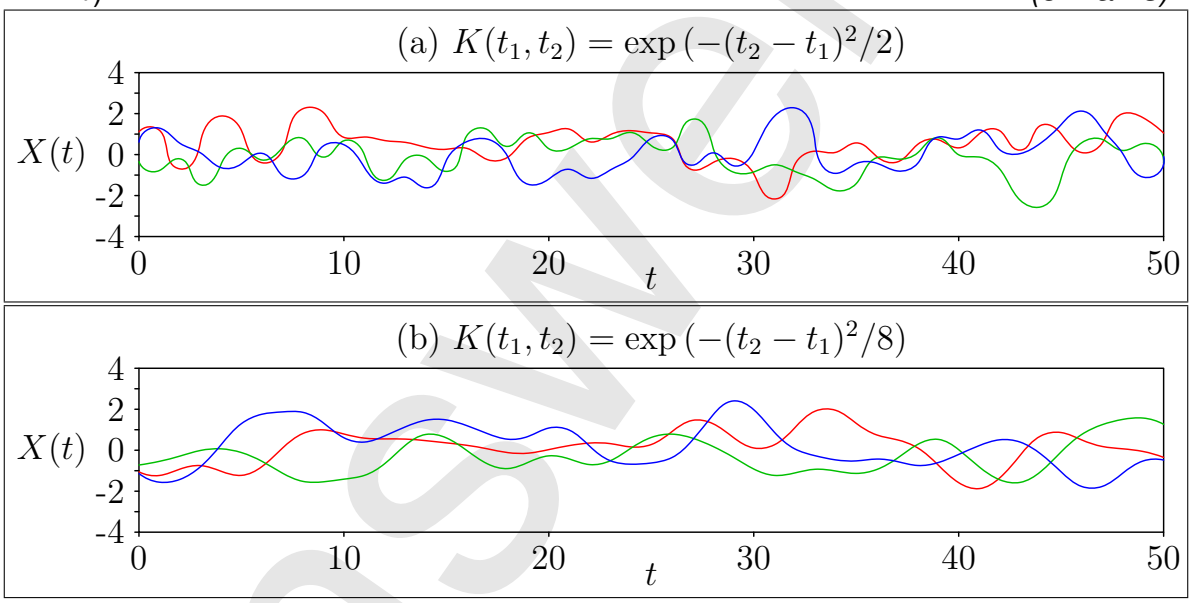
The kernel give the two point correlation function. If we have enough data point $(\boldsymbol{x}_i, y_i)$ we can compute the pairwise distances $d = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$. We can then compute the mean correlations $(y_i - \mu)(y_i - \mu)/\sigma^2$ for all pairs of points whose distance 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.

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)*



5

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

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).

5

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

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).

5

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

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).

5

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

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.

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