Wednesday, 26 April 2017
2.00 pm – 4.00 pm
(Duration: 2 hours)

DEGREES OF MSc, MSci, MEng, BEng, BSc,MA and MA (Social Sciences)

Machine Learning M

(Answer 3 questions in Section A and all of Section B)

This examination paper is worth a total of 60 marks

The use of a calculator is not permitted in this examination

INSTRUCTIONS TO INVIGILATORS

Please collect all exam question papers and exam answer scripts and retain for school to collect. Candidates must not remove exam question papers.

Section A

1. Linear regression with models of the form $t_n = \sum_{k=0}^{K} w_k \, h_k(\mathbf{x}_n) = f(\mathbf{w}, \mathbf{x}_n)$, is a common technique for learning real-valued functions from data.

(a) Squared and absolute loss are defined as follows:

$$L_{squared} = (t_n - f(\mathbf{x}_n, \mathbf{w}))^2, \quad L_{absolute} = |t_n - f(\mathbf{x}_n, \mathbf{w})|$$

Describe, with a diagram if you like, why, when optimizing the parameters with the squared loss, outliers have a larger effect than with the absolute loss.

[3]

(b) Which one of the following statements is true:

A) Parameter estimation with the squared loss is not analytically tractable.

B) The use of squared loss is equivalent to assuming normally distributed noise.

C) Absolute loss is a popular prior in Bayesian inference.

[1]

(c) Discuss why the value of the squared loss on the training data cannot be used to choose the model complexity.

[3]

(d) For the particular model $t_n = w_1 x_n + w_2 x_n^3$, I optimize the parameters and end up with $\mathbf{w} = [2,1]^T$. What does the model predict for a test point at $x_{new} = 3$?

[1]

(e) When performing Bayesian analysis in the regression model, we must choose a *prior* distribution over the parameters and a *likelihood* function. Describe the role of these two objects.

[4]

(f) Which of the following statements is the correct definition of a conjugate prior and likelihood pair?

A) The prior and likelihood are conjugate if they are of the same form (e.g. both are Gaussian).

B) The prior and likelihood are conjugate if they result in a posterior of the same form as the prior.

[1]

(g) What is the implication for parameter inference when the prior and likelihood do not form a conjugate pair?

[2]

2. Evaluating the predictive performance of trained methods is a vital task in the application of machine learning.

(a) Using a machine learning problem of your choice (e.g. regression, classification, etc). Describe what is meant by:

(i) Generalisation

[2]

(ii) Over-fitting

[2]

(b) A classification algorithm has been used to make predictions on a test set, resulting in the following confusion matrix:

	Truth			
Prediction		**Positive**	**Negative**	**Total**
	Positive	23	5	28
	Negative	10	12	22
	Total	33	17	50

Compute the following quantities (expressing them as fractions is fine):

(i) Accuracy

[1]

(ii) Sensitivity

[1]

(iii) Specificity

[1]

(c) Explain why it is not possible to compute the AUC from a confusion matrix.

[2]

(d) Three binary classifiers are used to make predictions for the same set of four test points. These predictions are given below, along with the true labels. In each case, compute the AUC (hint: you don't need to plot the ROC curve).

Classifier 1		Classifier 2		Classifier 3	
Predicted probability of class 1	True class	Predicted probability of class 1	True class	Predicted probability of class 1	True class
0.9	1	0.9	1	0.9	1
0.8	1	0.8	0	0.8	0
0.3	0	0.3	0	0.3	1
0.2	0	0.2	1	0.2	0

[3]

(e) Describe what is meant by cross-validation, including motivation for using it in model validation.

[3]

3. Classification is one of the most common problems in machine learning.

(a) Describe the key difference between classification and clustering.

[1]

(b) Describe what is meant by support vectors in the Support Vector Machine (SVM).

[2]

(c) In the context of SVMs, when and how should the kernel be used?

[2]

(d) Describe how the decision boundary is defined in logistic regression and give one advantage it has over the SVM's boundary.

[2]

(e) Let $p(\mathbf{w}|\mathbf{t}, \mathbf{X})$ be the posterior distribution of the parameters in a logistic regression model.

(i) Outline how to use Laplace approximation to approximate $p(\mathbf{w}|\mathbf{t}, \mathbf{X})$

[4]

(ii) The Metropolis-Hastings algorithm can also be used to infer $p(\mathbf{w}|\mathbf{t}, \mathbf{X})$. Provide pseudo code for the Metropolis-Hastings algorithm.

[4]

4. K-means and Gaussian mixture models are two popular clustering methods.

(a) Describe an application of clustering.

[1]

(b) Provide pseudo code for K-means (assume that the number of clusters is provided).

[4]

(c) Describe the key differences between K-means and Gaussian mixture models.

[4]

(d) Given that the squared Euclidean distance between data point $\mathrm{x_n}$ and the kth cluster center $\mathrm{\mu_k}$ is $(\mathrm{x_n} - \mathrm{\mu_k})^{\mathrm{T}}(\mathrm{x_n} - \mathrm{\mu_k})$, and $\mu_k = \frac{1}{n_k}\sum_{i=1}^{N} z_{ik}\,\mathbf{x}_i$, show that the kernelised equivalent is the following:

$$\mathrm{K(x_n, x_n)} - \frac{2}{\mathrm{n_k}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{z_{ik}\,K(x_n, x_i)} + \frac{1}{\mathrm{n_k^2}}\sum_{\mathrm{i=1}}^{\mathrm{N}}\sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{z_{ik}\,z_{jk} K(x_i, x_j)}$$

where n_k is the number of objects assigned to μ_k, and z_{nk} is 1 if object n is assigned to cluster k and 0 otherwise.

[4]

(e) Outline a strategy for selecting the number of clusters.

[2]

Section B

5. This question is based on 'Bayesian Classification with Gaussian Processes', Williams and Barber, IEEE PAMI 20(12):1342—1351

(a) Briefly summarise the contribution of this paper.

[2]

(b) The authors are interested in computing the classification probability $P(c|\boldsymbol{x})$**,** where c is the class and $\mathbf{x}$ is some observation. To do this they define a function $y(\boldsymbol{x})$ that outputs a real value and then push this value through the following sigmoidal function:

$$P(c|\boldsymbol{x}) = \frac{1}{1+e^{-y(x)}}$$

Why is this sigmoidal function required?

[2]

(c) The authors use a Gaussian Process (GP) for the function $y(\boldsymbol{x})$. Briefly describe what a Gaussian process is, and the advantage it has over a model of the form: $y(\boldsymbol{x}) = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}$.

[4]

(d) The authors begin by discussing Gaussian Processes for regression before moving onto classification. Computationally the regression problem is easier than classification with GPs. Why?

[2]

(e) The Gaussian Process uses a covariance matrix (which is very similar to a kernel matrix in an SVM). What do the authors propose for optimizing these parameters?

[2]

(f) The authors provide some experiments to evaluate their approach. Provide a critique of their experiments. For example, what do you think they did well? What did they not do well? Is there anything they omit that you feel would strengthen this section?

[3]