University of Glasgow

Friday 27 April 2018
9.30 am – 11.30 am
(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 $t_n = \sum_{d=1}^{D} w_d \, x_{n,d}$, $n = 1, \ldots, N$, is a common technique for learning real-valued functions from data.

(a) When estimating the parameters using the average squared loss function, it is common to regularise the parameters with a penalty term $\lambda \sum_{d=1}^{D} w_d^2, \lambda > 0$. Write down the combined loss function in vector/matrix form and give clear definitions of each variable: what is its dimension, and what do the elements represent.

[2 marks]

(b) Show that the optimal solution to **w,** which minimises the loss in **(a)**, is $\mathbf{w} = (\mathbf{X}^T\mathbf{X} + N\lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{t}$

[4 marks]

(c) Discuss the effect of the regularised loss function in **(a)** on the estimated parameters.

[2 marks]

(d) The L1 norm of the parameter, $\sum_{d=1}^{D} |w_d|$, is also a common way to regularise the loss function. Explain the effect it has on parameter estimation. Give an example of when it should be applied.

[2 marks]

(e) A powerful way to generalize the linear model is to apply basis function to the input variable. For example, the polynomial functions. Please describe a C-fold cross validation scheme to determine the appropriate order of the polynomial terms

[5 marks]

2. Bayesian inference is an important tool in machine learning.

(a) Describe the difference between Bayesian inference and maximum likelihood approach when estimating unknown parameters.

[2 marks]

(b) Let $p(\mathbf{w}|\mathbf{t}, \mathbf{X})$ be the posterior distribution which can be written as

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \frac{p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

The term on the denominator is known as the *marginal likelihood*. Write it down in its full form

[1 mark]

(c) If the likelihood and prior are from a conjugate family, which of the following is not correct?

A. $p(\mathbf{t}|\mathbf{X})$ will be analytically tractable

B. $p(\mathbf{w}|\mathbf{t},\mathbf{X})$ will always be of the same form as $p(\mathbf{w})$

C. $p(\mathbf{w}|\mathbf{t},\mathbf{X})$ will always be of the same form as $p(\mathbf{t}|\mathbf{X},\mathbf{w})$

D. $p(\mathbf{w}|\mathbf{t},\mathbf{X})$ will be of the same form as $p(\mathbf{t}|\mathbf{X},\mathbf{w})$ in certain parametric families (e.g. Gaussian)

[2 marks]

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

[4 marks]

(e) Describe a situation where the Metropolis-Hasting algorithm might perform badly.

[2 marks]

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

[3 marks]

(g) Please state the key assumption of the Laplace approximation.

[1 mark]

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

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

[2 marks]

(b) What is the key assumption in a Naive Bayes classifier?

A. That the problem is binary (there are only two classes)

B. The features are independent

C. The data are Gaussian distributed

D. The features are independent conditioned on a particular class

[1 mark]

(c) Explain what overfitting is in the context of classification (you might find a graph useful)

[2 marks]

(d) How to plot the ROC curve of a binary classifier (state the meaning of x-axis and y-axis in the plot).

[3 marks]

(e) Two 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	
Predicted probability of class 1	True class	Predicted probability of class 1	True class
1	1	1	1
0.7	1	0.9	1
0.6	1	0.6	0
0.5	0	0.5	1
0.4	0	0.2	0
0.0	0	0.0	0

[2 marks]

(f) Which of the following statements about the Area Under the ROC curve (AUC) (in the context of binary classification) is not true:

A. AUC is unreliable when class sizes are imbalanced.

B. The highest possible value of AUC is 1

C. If a classifier achieves an AUC less than 0.5 it will perform better if all of its predictions are reversed

D. AUC and ROC can't be directly applied to multiclass cases.

[1 mark]

(g) Explain how the SVM can be extended via the kernel trick to perform non-linear classifi- cation.

[2 marks]

(h) What is the advantage of performing non-linear classification with the SVM?

[1 mark]

(i) We train a logistic regression model (with sigmoidal function, $1/(1+\exp(-\mathrm{w}^{\mathrm{T}}\mathrm{x}))$) resulting in a parameter vector of $\mathrm{w} = [1,-1,2,2]^{\mathrm{T}}$. What is the output probability for the input vector $\mathrm{x} = [4,2,1,-2]^{\mathrm{T}}$?

[1 mark]

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

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

[4 marks]

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

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

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

[4 marks]

(c) Is the total Euclidean distance between data points and their cluster centers a good criterion to select number of clusters in K-means? Why?

[2 marks]

(d) Gaussian mixture models can be fitted to data using the expectation maximization (EM) algorithm. The EM algorithm has two steps: E-step and M-step. Describe what parameters are being estimated in each step in Gaussian mixture models.

[2 marks]

(e) Describe three key differences between K-means and Gaussian mixture models

[3 marks]

Section B

5. This question is based on '*The Infinite Gaussian Mixture Model. Carl Rasmussen. NIPS '12*

(a) Briefly summarise the contribution of this paper.

[2 marks]

(b) What is the benefit of having an infinite number of components?

[2 marks]

(c) What method did the author use for inferring the parameters? Explain the key steps in this method.

[4 marks]

(d) What is the model output for number of mixing components?

[2 marks]

(e) In the finite case (introduced in the paper), what prior distribution are used for i) the mean of each Gaussian component; ii) the precision of each Gaussian component; and iii) the mixing coefficients.

[2 marks]

(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 marks]