University of Glasgow

Tuesday 17 December 2019
9.30 am – 10.30 am
(Duration: 60 minutes)

DEGREE OF MSc

Machine Learning for Data Scientists

(Answer all of the 3 questions)

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.

1. A polynomial regression model is defined as:

$$t_n = \sum_{d=0}^{D} w_d \, x_n^d, n = 1, \ldots, N$$

(a) When applying this model to the Olympic data, where $x_n \in 1896, \ldots, 2008$, we always rescale x, e.g. $x_n = \frac{x_n - 1896}{40}$. Explain why this rescaling is necessary.

[4 marks]

(b) Write down the regression model if the following Radial basis function (RBF) with a range of different position parameter μ is applied to x_n.

$$RBF(y; \mu, l) = \exp(-\frac{(y-\mu)^2}{l^2})$$

[3 marks]

(c) Give an example of nonlinear regression model. Also explain why it is nonlinear.

[2 marks]

(d) Explain why polynomial regression could suffer from outliers.

[3 marks]

(e) L2 Regularised regression can be used to deal with this problem. Use a contour plot (Assuming the dimension of the parameter is 2) of parameters and loss function to explain why.

[8 marks]

2. Classification question

(a) Classification and regression are both supervised learning problems. Describe a way to turn a regression problem into a classification problem? (Please state the differences between the two.)

[3 marks]

(b) The receiver operating characteristic (ROC) curve is a standard way to visualize the performance of classifiers. Outline how to draw a ROC curve

[3 marks]

(c) For the classifier outputs in the table below, provide a value for the missing output (labeled '?') that would:

Class Label	0	0	0	1	1	1
Output	0.1	0.25	0.4	?	0.6	0.9

(i) Give an AUC equal to 1,

[2 marks]

(ii) Give an AUC less than 1,

[2 marks]

(iii) Now assuming you can change any output of the six data points, give an example of the output, such that the AUC is 0.5.

[3 marks]

(d) Use a diagram (some data in 2D) to describe how linear Support Vector Machines (SVMs) operate (how they make classification decisions, what and how parameters have to be set, what data needs to be stored etc).

[4 marks]

(e) Logistic regression uses the sigmoid function to make classification decision. Now, we have defined the following probability with a sigmoid function.

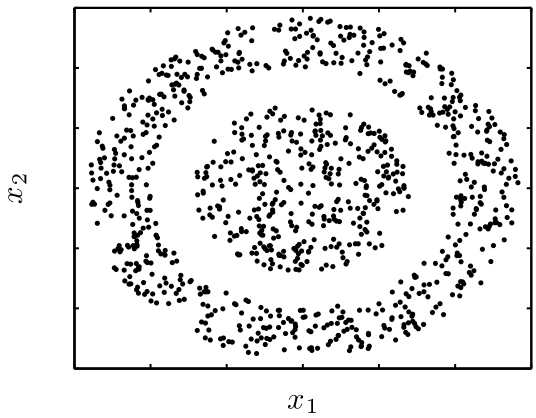
$$p(t_n = 0|\mathbf{w}, \mathbf{x}_n) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x}_n)}$$

Write down the corresponding likelihood function for n data points, $p(t_n|\mathbf{w}, \mathbf{x}_n)$.

[3 marks]

3. Unsupervised learning question

Consider using the K-means algorithm to perform clustering on the following data.



We want to cluster data in the outer ring in one cluster and the data in the inner circle as a different cluster.

(a) Outline what would happen if we directly apply *K*-means with Euclidian distance to this data. Can it achieve the clustering objective? How will it split/group the data?

[2 marks]

(b) An alternative approach is to use *Kernel K-means.* Explain how kernel could help in this dataset.

[3 marks]

(c) Which one of the following statements about kernel is NOT correct?

A. One could use the RBF kernel, $K(\boldsymbol{x}_n, \boldsymbol{x}_i) = \exp(-\gamma(\boldsymbol{x}_n - \boldsymbol{x}_i)^T(\boldsymbol{x}_n - \boldsymbol{x}_i))$, to achieve the clustering target.

B. The RBF kernel projects the data onto an infinite dimensional space. The free parameter γ can be estimated with cross-validation.

C. One could use the linear kernel, $K(\boldsymbol{x}_n, \boldsymbol{x}_i) = \boldsymbol{x}_n^T\boldsymbol{x}_i$, to achieve the clustering target.

D. The linear kernel projects the data onto itself, and there is no free parameter to tune.

[2 marks]

(d) Write some pseudo-code to perform K-means.

[5 marks]

(e) Outline how and why cross-validation can be used to select the number of clusters K.

[8 marks]