University of Glasgow

DayOfWeek DayOfMonth Month 2XXX
XX.XX am/pm – XX.XX am/pm
(Duration: 60 minutes)

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

Machine Learning for Data Scientists

(Answer all of the 3 questions)

This examination paper is worth a total of 60 marks

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 is this rescaling necessary.

[4 marks]

In the Olympic data, high polynomial could make the value of x^d very big [2]. As a result, the computation of parameters becomes unstable [2].

(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]

$$t_n = \sum_{d=1}^{D} w_d \exp(-\frac{(x_n - \mu_d)^2}{l^2}), n = 1, \ldots, N$$

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

[2 marks]

Any nonlinear function of the w and x e.g. f(x, w) = sin(w*x) [1]. State clearly which relationship is nonlinear [1].

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

[3 marks]

Outliers favours higher polynomial orders [1], as they can fit the outliers better [1]. However, the high order models are more likely to overfit the data [1].

(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 the loss function to explain why.

[8 marks]

2 marks for the correct contour of the mean squared error, 2 marks for the correct contour for the L2 regularisation. 2 marks for highlight the correct intersection between the two. 2 marks for stating the fact that optimal parameter shifted closer to (0, 0)

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]

In regression, the target variable is a continuous/real-valued variable [1]. In classification, the target variable is a discrete/binary/categorical variable [1]. To obtain a classification problem from regression, one can cluster/group the continuous/real-valued variable into groups [1].

(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]

The ROC curve is created by varying the threshold at which the classifier calls something as belonging to the positive class. [1] Each point is consisted of false positive rate or 1-specifity (normally on the x-axis) [1] and true positive rate or sensitivity (normally on the y-axis) [1].

(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]

Anything between 0.4 and 1.0

(ii) Give an AUC less than 1,

[2 marks]

Anything less than 0.4

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

[3 marks]

As long as the output for 0s and 1s are the same.

(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]

1 mark for drawing that demonstrate the margin, 1 mark for highlight the decision boundary, 1 mark for highlighting support vectors, 1 mark for pointing out only support vectors needs to be stored.

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

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

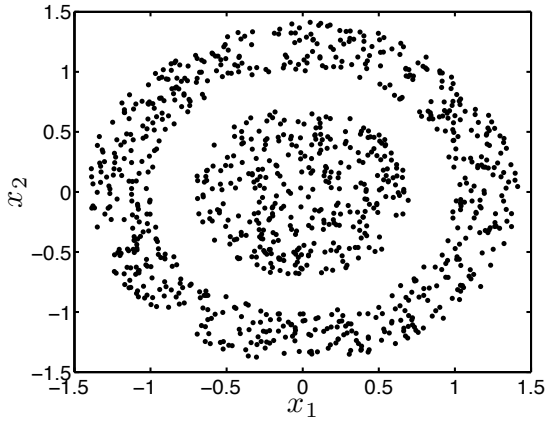
Write down the corresponding likelihood function for nth data points, $p(t_n|\mathbf{w}, \mathbf{x}_n)$.

[3 marks]

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

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]

K-means cannot split the data into outer ring and inner circle clusters [1]. It will group parts of the outer ring or inner circle [1].

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

[3 marks]

A kernel that project the data onto a different space [1] where data can be easily separated [1]. The space could have higher or lower number of dimensions compare to the original data [1].

(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]

C

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

[5 marks]

Given: Number of clusters, K

2. For each cluster k = 1...K:

3. For each object n = 1...N: [1]

4. Compute the distance between object n and cluster k [1]

5. Assign object n to the cluster corresponding to the smallest distance [1]

6. Update the mean of each cluster [1]

7. If assignments have changed, return to 2. Else stop. [1]

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

[8 marks]

Cross-validation with total or average Euclidean distance between data points and their cluster centers [2]. At each CV cycle, use the training data to determine the mean of the clusters [2]. Test these means by computing total or average Euclidean distance between testing data points and their nearest cluster centers [2]. CV allow the trained number of clusters and means to be tested on a different dataset in which the trained K-mean may or may not be a good fit. [2]