University of Glasgow

DayOfWeek DayOfMonth Month 2XXX
XX.XX am/pm – XX.XX am/pm
(Duration: 90 minutes)

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

Machine Learning & Artificial Intelligence 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. Considering linear regression on the Olympic data in figure 1.

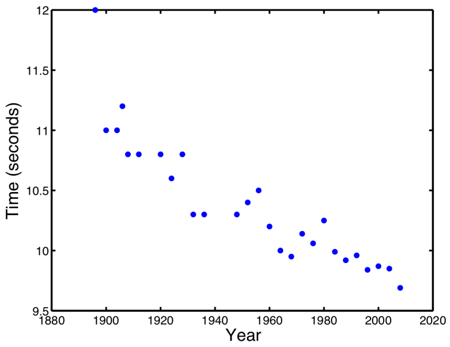

Figure 1: Olympic data

(a) We want to predict Olympic years from 100m winning times. What should be the target value and attribute? When solving this regression task with a polynomial regression model, how would you rescale the attributes? Why?

[6 marks]

Target: years [1], attribute: winning times [1]. A reasonable solution with sufficient details. E.g. whiting (x-mean(x))/std(x) [2]. A reasonable explanation of what the solution can do. E.g. Whiting makes sure the attribute is in [2].

(b) Based on what you have learned from fitting linear regression models (with polynomial or RBF) to the relationship between years and winning times, predict which year may produce winning time 9s and 13s, explain why.

[4 marks]

Answer should include reasonable estimation of years that may produce winning time 9s and 12s [2], using arguments from existing data and model, could be polynomial or RBF [2].

For example, polynomial order of 3 might be a good fit to the data, the model is likely to predict 9s after 2040, 13s could be before 1860.

(c) The radial basis function (RBF):

$$h_{n,k} = \exp\left(-\frac{\sum_{d=1}^{D}(x_{n,d} - \mu_{d,k})^2}{2s^2}\right), n = 1, \dots, N;\ k = 1, \dots, K$$

is a popular basis function. The parameter $\mu_{d,k}$ is often be a data point $x_{i,d}, i = 1, \dots, N$. Outline the strength and risk of this setup for $\mu_{d,k}$, and how would you mitigate the risk.

[5 marks]

Strength: flexibility [1]. Risk: numerical stability and overfitting [2]. Use less centers or add small value to the diagonal of X^TX, using regularization [2].

(d) In addition to the polynomial function and RBF, linear regression can be generalized using other basis functions. One of most widely used example is the Fourier analysis, let's consider the following linear regression model:

$$t_n = \sum_{j}^{m} A_j \cos(jx_n + \theta_j)$$

What is the basis function of choice here? How would you deal with the unknow parameters A_j and θ_j? (Hint: you might find the following trigonometry identity useful, $\cos(a+b) = \cos(a)\cos(b) + \sin(a)\sin(b)$).

[5 marks]

Solution 1[3 marks in total]: Basis function: for $\cos(jx_n + \theta_j)$[1], A_j is the regression parameter, cross validation for \theta [2]

Solution 2 [full mark]: Two basis functions as a result of applying the provided identity $\cos(jx_n)$ and $\sin(jx_n)$. \theta_j becomes part of the linear regression parameter, the same as A_j.

2. Classification question

(a) The likelihood of logistic regression

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

where $g(a) = \frac{1}{1+\exp(-a)}$. Use an example of a few data points to explain how the likelihood function tells how well the parameter **w** fits the data.

[4 marks]

The example needs to have more than one pair of t_n and x_n [1], demonstrating how to construct joint data likelihood [1]. It should also include a parameter estimate representing a good fit and a parameter estimate representing a bad fit [2].

(b) The following matrix contains estimated parameters values from three types of logistic regression models. The model type is indicated by the columns. The parameter of each feature is placed in the corresponding row. Give your best estimate of what each model is and explain why.

```
           Model 1          Model 2           Model 3
[[  1.08381535e+01   1.19648635e+01   1.11285803e+01]
 [-0.00000000e+00  -1.29443055e+01  -3.29359603e-01]
 [-0.00000000e+00   5.79522897e+01  -1.94725736e-01]
 [-1.16126582e-01  -1.09582035e+02  -9.64898104e-02]
 [-1.59001968e-02   6.23248849e+01  -1.87327081e-02]
 [-0.00000000e+00   7.48519704e+01   3.32402164e-02]
 [ 1.38119952e-03  -1.46955431e+02   4.50182751e-02]
 [ 3.22128802e-03   1.04735797e+02   9.53751777e-03]
 [ 1.61616847e-04  -3.88035781e+01  -3.60588365e-02]
 [-8.65262203e-05   7.43343695e+00   1.40369595e-02]
 [-7.74413350e-05  -5.82870289e-01  -1.62830483e-03]]
```

[6 marks]

Model 2 [2]: logistic polynomial regression. Some parameter values are very big in absolute value. Model 3 [2]: L2-regularised logistic regression. Compare to model 1, most parameters are much smaller in absolute value. Model 1 [2]: L1-regularised logistic regression. Some parameters are exactly zeros.

(c) Compare the effect on prediction of the three logistic models in **(b)**.

[4 marks]

With the same x_n [1], L1- and L2-regularised logistic regression are likely to produce lower probability of being the positive class [2], they have better generality with unseen data [1].

(d) Let's consider a binary classifier trained on a falsely labeled dataset. The issue is all positive (*1*) and negative (*0*) labels are swapped during training. The classifier outputs in the table below:

Correct label	0	0	0	1	1	1
False label during training	1	1	1	0	0	0
Probability of being the *positive* class	0.9	0.8	0.7	?	0.2	0.1

(i) What would be the AUC (computed with the correct labels) when the classifier is perfectly trained on the false data? And why?

[2 marks]

0, Perfect AUC on false label is 1. Correct label is 1-1. Or arguing using the definition of AUC: the probability of a positive example having classification score higher than a negative one.

(ii) Provide the range of possible values for the missing output (labeled '?') that would be produced by the classifier in (i). Explain why.

[2 marks]

[0, 0.7), all number ensures that all positive data (based on false label) have high score than any negative data (based on false label).

(iii) What would be the AUC (computed with the correct labels) of a random classifier trained on the falsely labeled data? Why?

[2 marks]

0.5, random classifier will have 0.5 AUC on false label, 1-0.5 is still 0.5.

3. Clustering question (Figures in this question were taken from the sklearn clustering tutorial: https://scikit-learn.org/stable/modules/clustering.html)

(a) Describe clustering results of K-means and Gaussian Mixture in figure 2. Hint: answer should address parameters estimation, initial conditions and selecting the number of clusters.

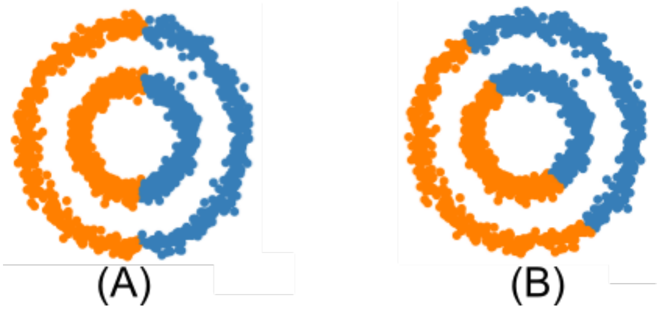

Figure 2: Clustering results of (A) K-means and (B) Gaussian mixture model.

[6 marks]

3 marks each: answer should address parameters estimation [1], initial conditions [1] and selecting the number of clusters [1].

Example solution:

Kmeans: two cluster centers are located somewhere between the two rings (closer to the inner ring) [1]. It's hard to determine what initial conditional might have been used. Many splits are possible. [1] Difficult to do model selection for the same reason (Many splits are equally well) [1].

GMM: the means are also located somewhere between the two rings. The two covariance matrices and mixing weights should be almost identical between the two [1]. Similar to Kmeans, there are many equally good solutions for GMM. Specific initial condition is hard to determine [1]. Similar to Kmeans, model selection is hard due to multiple good solutions in each number of cluster setting [1].

(b) Suppose we want to avoid any data point from the inner ring being assigned to the same cluster with any point data point from the outer ring. Outline two approaches to achieve this goal with the Gaussian mixture model? Hint: You don't have to use just 2 clusters.

[4 marks]

Project the data onto to a different space [1] where data points in the inner ring are well separated from the data points in the outer ring [1]. Restrict variance of each component to be small [1] and use more clusters [1].

Marks for any other sensible approaches

(c) Describe clustering results of K-means and Gaussian Mixture in figure 3. Hint: answer should address parameters estimation, initial conditions and selecting the number of clusters.

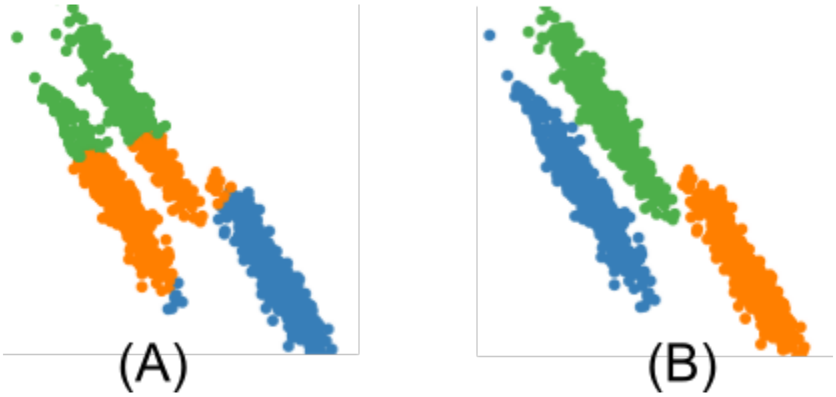

Figure 3: Clustering results of (A) K-means and (B) Gaussian mixture model.

[6 marks]

3 marks each: answer should address parameters estimation [1], initial conditions [1] and selecting the number of clusters [1].

Example:

Kmeans: three cluster centers are vertically distributed [1]. Initial cluster centers may also have been vertically distributed while others are possible [1] A cross validation on clustering number might prefer the number of clusters larger than 3. [1].

GMM: three highly Gaussian-like distributed data. Different means but similar covariance matrices and mixing weights [1]. In this case, the optimal solution is quite clear, therefore, GMM should be same with multiple initial with only the identity the clusters changing [1]. Cross-validation or Marginal likelihood should favor a 3 cluster solution [1].

(d) Suppose Figure 3 (B) represents the results we want. Outline one approach to achieve this goal with K-means. Hint: Sufficient details of the approach are required to get full marks.

[4 marks]

Kernel k-means [2]. Details on how to kernelize K-means: using the kernel trick on the distance [1] and only update the assignments [1].