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.

Question 1: Regression (Total marks: 20)

Consider using regression to predict the birth rate in the US using the data shown in the following figure:

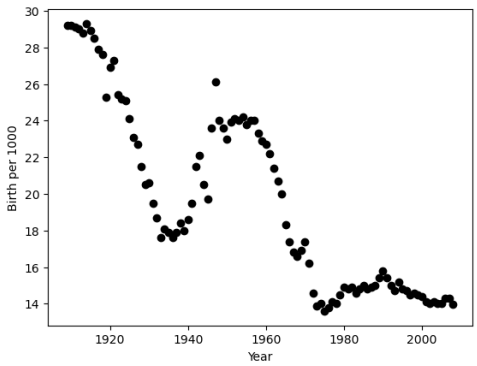


Figure 1.1 Birth rate (per 1000) from 1909 to 2008

(a) Consider fitting the data with a polynomial regression of order 10. Identify the numerical issue with model fitting and propose a solution with sufficient details

[4 marks]

The value of the year could cause numerical issues when raised to order 10 [1], the numerical issue is matrix inversion in the least square solution being unstable when year^10 is too large [1]. 2 marks for a reasonable rescaling method as a solution, including whitening, min-max, or take the logarithm.

(b) Consider fitting the data with a polynomial regression of order 2, identify the two regions of most likely poorly fitted data points and explain why.

[6 marks]

2 marks for identifying the correct poorly fitted data points, several options: (x =1945-1960, y ~=24), (x ~= 1938-1940, y ~=17), Data points with x range from 1975 to 1980.

4 marks for reasoning: A polynomial regression model with an order of 2 is a quadratic (or convex) curve [1]. The global minima of the curve would be in years after 2008[1]. The left tail end would be following the downward trend from 1909 to 1920. [1]. The curve would cut through the drop from 1920-1940 and the upward trend after the 1940s.[1]. Option mark {1}: The fitted curve will struggle to capture the drop in 1970-1980.

(c) Consider fitting the data in Figure 1.1 with a liner regression model with the sigmoid basis function:

$$h_{n,k} = sigmoid\left(\frac{(x_n - \mu_k)^2}{s}\right), n = 1, \dots, N;\ k = 1, \dots, K.$$

Explain the choice of hyperparameter μ_k (mu_k) and s that could lead to the following fitted model

[4 marks]

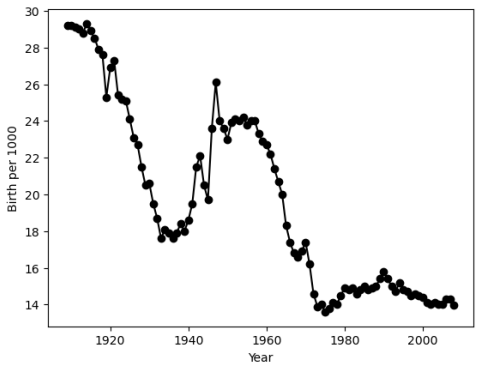


Figure 1.2 A liner regression model using sigmoid basis function fitted to the data

Mu_k can be set as x_n [1], and s is set to relatively small in relation to x_n, for example, 1 [1]. Setting the centre parameter to be x_n gives the model maximum flexibility to fit every data point [1]. Setting s to be small allows sharp turns between data points to be fitted [1].

(d) We used two fitting strategies, namely ridge regression and lasso, and obtained the following fitting models in Figure 1.3 A and B. Identify which fitting strategy is used in each figure and explain why and how the chosen fitting method could have generated the result. (note, each method is used only once).

[6 marks]

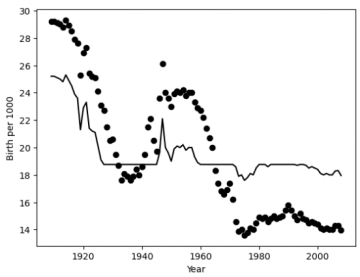


Figure 1.3 A

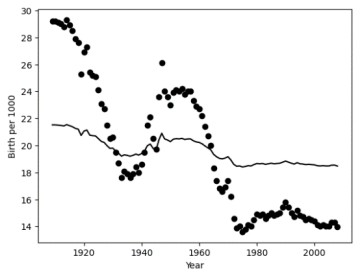


Figure 1.3 B

Figure 1.3 A is fitted with Lasso [1]. Reasoning: The fitted model has three flat regions. This happens when basis functions covering these regions have zero coefficients [1]. The fitted model also suggests that the hyperparameter controlling the strength of l1 regularization is set to be too strong [1, identifying l1 regularization is needed].

Figure 1.3B is fitted with ridge regression [1]. Reasoning: The fitted model is poorly following the trend. This happens when all basis functions have very small coefficients [1]. The fitted model suggests that the hyperparameter controlling the strength of l2 regularization is set to be too strong [1, identifying l2 regularization is needed].

Question 2: Classification (Total marks: 20)

a) Assume the following training data in the two-dimensional plane of X_1 and X_2 is available (Figure 1). The target variables for the points in the red and blue are +1 and -1. We summarise the data as the following tuples: <(2,0), 1>, <(0,2),-1>, <(0,-2),1>, and <(-2,0),1>, respectively.

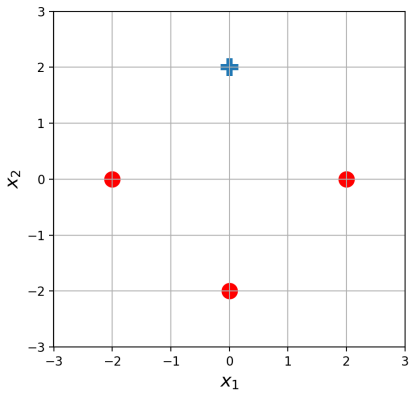


Figure 2

Deleted: *1*

i. Design a k-NN classifier with k=1 and use it to determine the class variables C_1 through C_4 for the following test data points: <(0,1), C_1>, <(1.5,1), C_2>, <(-0.5,1), C_3>, and <(0,0), C_4>:

[4 marks]

SOLUTION:
C_1 = -1, C_2 = 1, C_3 = 1, C_4=unknown

ii. What would be the class variable C_4, if we had used k=3? [2 marks]
SOLUTION:
C_4 = 1

iii. Write down the equations that specify the decision boundary between the two classes.
[4 marks]

SOLUTION:
X_1 – X_2 = 0 in X_1 > 0.0 and X_2 > 0.0 2 marks
X_1 + X_2 = 0 in X_1 > 0.0 and X_2 < 0.0 2 marks

b) In the same data set in Figure 1, we apply a linear SVM model with the predictor $y(X_1, X_2)$ for classification.

i. Which data points are the support vectors? Write down the equation for $y(X_1, X_2)$. (Hint: First visually assess the data to determine the decision boundary and the support vectors. Observe the constraints for the margin and SVM classifier.)
[6 marks]

SOLUTION:

Commented [JL1]: Should the sub-questions be numbered i,ii,iii,etc? Or for digital exams, is it better to flatten the whole question so there aren't sub-questions that don't show up nicely in Moodle?

Commented [AG2R1]: There is a shared information set in the preamble of the question. Flattening would ruin this and I insist on keeping the same structure as it is.

(2,0), (0,2), and (-2,0) 3 mark,
$y(X_1, X_2) = -2X_2 + 1$ 3 marks

ii. Specify the Lagrange multipliers $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ for each of the data points in the training data (2,0), (0,2), (-2,0), and (0,-2), respectively.

[4 marks]

SOLUTION:
$\alpha_1 = .5, \alpha_2 = 1, \alpha_3 = 0.5, \alpha_4 = 0$

Question 3: Unsupervised learning (Total marks 20)

Consider using the K-means algorithm to perform clustering on the following scenario Figure 3.1 A. We expect to form two clusters as shown in Figure 3.1 B.

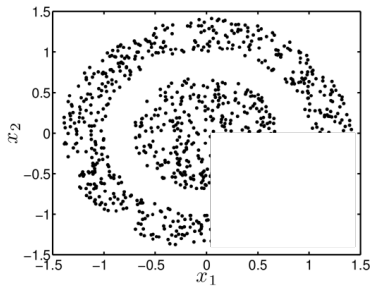


Figure 3. 3A: Original Data

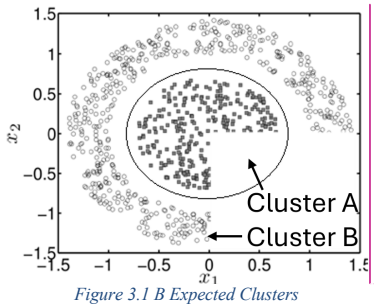


Figure 3.1 B Expected Clusters

Commented [JL3]: Figure might be better with (colorblind friendly) colors

Deleted: 2

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

[3 marks]

K-means clustering aims to partition the data into (k) clusters by minimizing the variance within each cluster. It does this by assigning each data point to the nearest cluster center (centroid) based on Euclidean distance (1 mark).

K-means assumes that clusters are spherical and equally sized, which is not the case with concentric circles. Concentric circles are non-spherical and have different radii.

Euclidean distance measures straight-line distance, which doesn't work well for circular or ring-shaped clusters. Points on the inner circle are closer to the centroid of the outer circle than to the centroid of their own circle. (1mark for identifying any of these issues)

K-means will likely split the data into arbitrary segments rather than correctly identifying the two concentric circles. For example, it might divide the data into pie-like slices or other non-intuitive shapes. (1mark)

(b) An alternative approach is to use *Kernel K-means.* Would kernel K-means could help in this dataset and why?

[2 marks]

Kernel K-means extends the traditional K-means algorithm by using a kernel function to transform the data into a higher-dimensional space where the clusters become more separable. (1 mark) A Radial basis function kernel would allow to handle the non-linear cluster boundaries. (1 mark)

(c) An alternative approach is to use *mixture models*. Would mixture models help to better classify this dataset than K-means and why?

[3 marks]

GMMs assume that the data is generated from a mixture of several Gaussian distributions with unknown parameters. (1 mark). GMMs assume that each cluster follows a Gaussian distribution, which is not ideal for ring-shaped data. (1 mark)

With points only on the boundaries, the model might place the means of the Gaussians in between the circles, leading to incorrect clustering. (1 mark)

(d) The plot in Figure 3.2 shows some 2D data. PCA is applied to this data. Explain how the first principal component would look if it is overlaid on the plot. Explain your reasoning. (Note: there is no need to make a drawing. You can provide a description of the shape based on the coordinate system provided in the original figure.)

Commented [JL4]: They won't be able to draw it, right? This is presumably an in-person digital exam

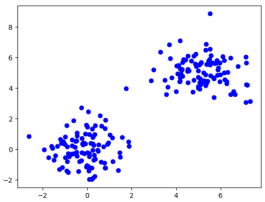
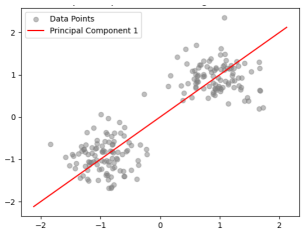


Figure 3.2 2D Points

[2 marks]

First principal component:



The first principal component will be across the direction of the highest variance (1 mark). In this case, we can note that there is a large variation in the axis that connect the two clusters (1 mark).

(e) Similar to the previous question, explain what the second principal component would look like and why. (Note: there is no need to make a drawing. You can provide a description of the shape based on the coordinate system provided in the original figure.)

Commented [JL5]: Again, I don't think they will be able to draw in the exam

[2 marks]

PCA decompose the data into orthogonal components (1 mark). Therefore, the second principal component will be orthogonal to the first: (1 mark)

(f) Describe the four-step process you should use to determine the number of clusters in Kernel K-Means. (Hint: Each step gets a mark.)

[4 marks]

1. Split the Data: Divide the data into (k) folds. (1 mark)
2. Kernel Computation: Compute the kernel matrix for the entire dataset. (1 mark)
3. Cross-Validation Loop: For each fold, train the Kernel K-means on (k-1) folds and validate on the remaining fold. (1 mark)
4. Evaluate Performance: Use a performance metric (e.g., silhouette score) to evaluate the clustering quality and average the results (1 mark)

(g) Describe two approaches you could take to managing the curse of dimensionality in, for example, genetic data. For example, how would you overcome this if you had a high-dimensional dataset with thousands of genetic features but only hundreds of subjects?

[4 marks]

Feature selection can help overcome the problem of the curse of dimensionality. (2 marks)

We can use a feature selection method that picks the most relevant features (1 mark), or we can use a dimensionality reduction technique like PCA that creates new features (1 mark)