Wednesday 7 December 2022
09:15-10:45 GMT
(Duration: 1 hour 30 minutes)
Additional time: 30 minutes
Timed exam – fixed start time

DEGREES OF MSc

Machine Learning & Artificial Intelligence for Data Scientists

COMPSCI5100

(Answer all 3 questions)

This examination paper is an open book, online assessment and is worth a total of 60 marks.

Question 1: Regression (Total marks: 20)

Consider using regression to predict the world population growth rate using the data shown in the following figure:

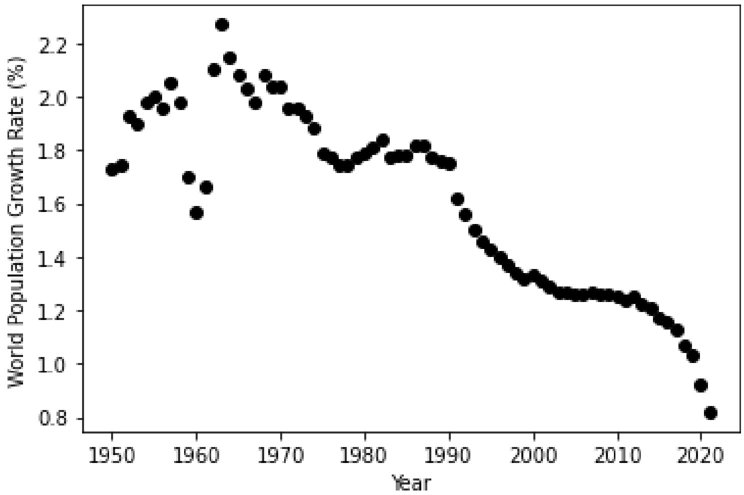


Figure 1. World population growth rate from 1950 to 2021. Source: https://ourworldindata.org/world-population-update-2022

(a) Propose a rescaling strategy (with enough details of the procedure) for the variable Year. Explain why the proposed strategy is appropriate.

[4 marks]

(b) Consider fitting the data with a polynomial regression model with the order of 1, identify the two most likely poorly fitted data points and explain why.

[6 marks]

(c) Consider fitting the data in figure 1 with a regression with the sigmoid basis function:

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

where x_n represents each year and *sigmoid*(a) = 1/(1+exp(-a)). Outline one advantage and disadvantage of using this sigmoid basis function over polynomials.

[4 marks]

(d) Suppose we use the sigmoid basis function in (c), with μ_k set to be x_n and $s = 10$, to to fit the data. We used three fitting strategies, namely linear regression, ridge regression and lasso, and obtained the following fitting model in Figure 2 A, B and C. Identify which fitting strategy is used in each figure and explain why (note, each method is used only once).

[6 marks]

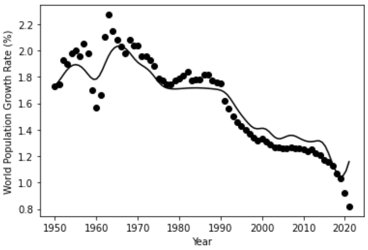


Figure 2 A

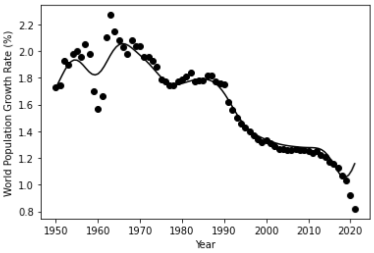


Figure 2 B

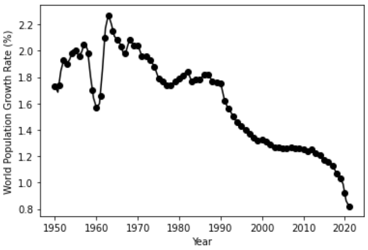


Figure 2 C

Question 2: Classification (Total marks: 20)

(a) You have been asked to design a classifier to automatically identify the Tweets that are considered as ‘hate speech’ in the social media website Twitter. You collected a training dataset which has 800 ‘regular’ tweets and 100 ‘hate’ tweets. Answer the following:

(i) Describe 2 features you might use for this task including their type (scalar/vector, real-valued or not).

[3 marks]

(ii) You learn a faulty classifier which always classifies a tweet as ‘regular’. What would be the weighted classification accuracy of this classifier?

[2 marks]

(iii) Assume that we use Logistic Regression for the task considering only 2 scalar features. We get the following set of parameters $\mathbf{w} = [-1.8, 2.1\ -0.3]^T$ after training. For feature vector $\mathbf{x} = [1,\ 1]^T$, calculate the output of the logistic function (probability score).

[4 marks]

(iv) Logistic Regression assumes a linear relationship between the dependent and the independent variables (x). Why is that considered a limitation of the model?

[2 marks]

(v) Can we replace the sigmoid function in Logistic Regression by the function g(z) shown in Fig. 3? Explain your answer.

[2 marks]

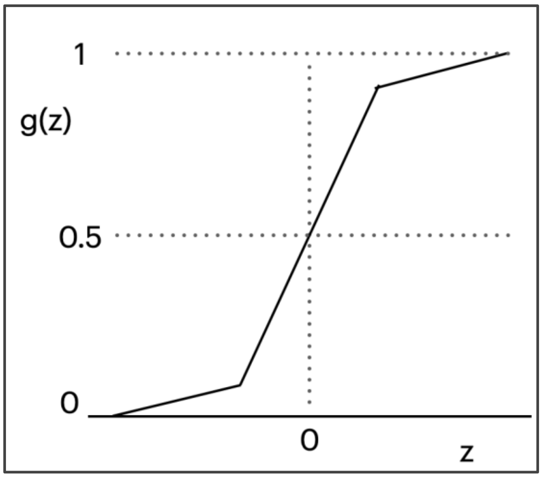


Figure 3: Proposed function

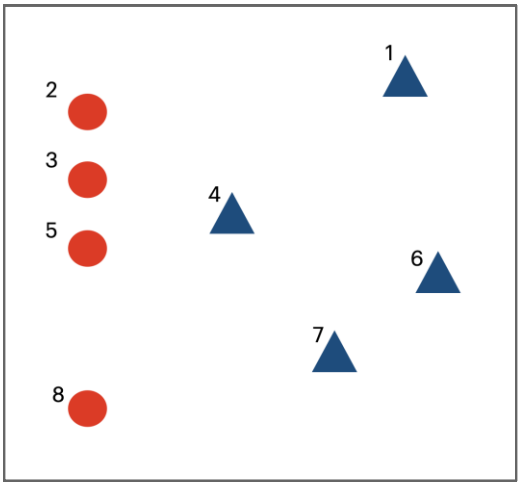


Figure 4: Training data points

(b) Consider the 8 data points for your training set belonging to 2 classes in Fig. 4. This is used to train a linear SVM.

(i) Draw the decision boundary for linear hard margin SVM method with a solid line. Show the margin using dotted lines.

[3 marks]

(ii) Which ones are the support vectors?

[2 marks]

(iii) What is the training error?

[1 marks]

(iv) Removal of which data point will change the decision boundary?

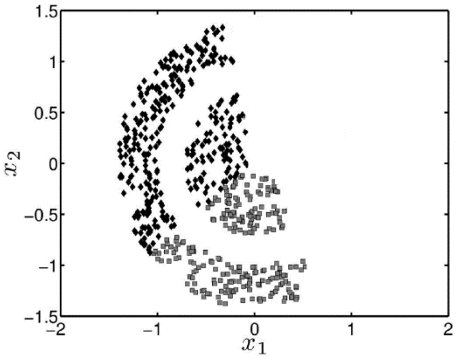
[1 marks]

3. Unsupervised learning question (Total marks 20)

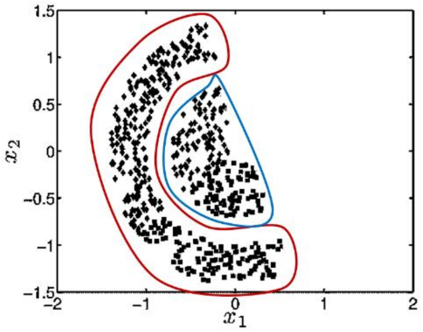
Consider using the K-means algorithm to perform clustering on the following scenario A1.

We expect to form two clusters as shown in A2.

A1) Original Data



A2) Expected clusters



(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?

[2 marks]

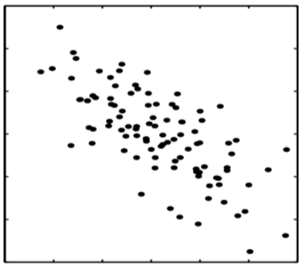
(b) An alternative approach is to use *Kernel K-means.* Would kernel K-means could help in this dataset and why?

[3 marks]

(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]

(d) The plot in D1 shows some 2D data. PCA is applied to this data. Sketch this plot, and indicate on your sketch what the first principal component would look like. Explain your reasoning.



D1) 2D points

[2 marks]

(e) Similar to previous question, sketch what the second principal component would look like and explain why.

[2 marks]

(f) Explain why PCA is used (provide at least two applications) and how to decide for the optimum number of principal components.

[2 marks]

(g) Explain the advantages and disadvantages of feature selection based on projection compared to feature selection based on how well they can discriminate between two classes.

[6 marks]