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 world population growth rate using the data shown in the following figure:

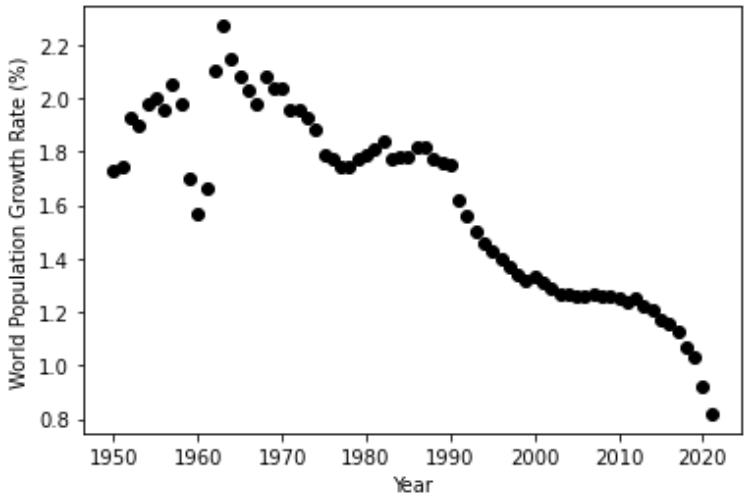


Figure 1. World population growth rate from 1950 to 2021. Source: https://ourworldindata.org/world-population-update-2022

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

[4 marks]

2 marks for a reasonable strategy, including whitening, min-max, or take logarithm. 2 marks for the reasoning, the key is to reduce the absolute value of "year", such that high order polynomial will still produce well-behaved values (small) [1] and the matrix inversion in the least square solution is still stable [1].

(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]

2 marks for identifying the correct poorly fitted data points, three options: x= (1960, ~1963, and 2021). 4 marks for reasoning: polynomial regression model with an order of 1 is a straight line [1], and the data in the figure can be split into two regions: before 1990 where growth rate changes a lot but stays the same on average [1], and after 1990 where growth rate consistently slows down [1]. A straight line needs to average over both regions and is therefore likely to miss the dramatic drop and jump around 1960 [1]. [Alternative answer to the final mark: or a quick drop in growth rate close to 2021.]

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

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

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]

Advantage: the variance is not equally distributed across x values, larger in small values and small in large x values [1]. Using location-specific basis functions sigmoid can model this localised effect better than polynomial functions which model global effect across all x values, resulting in better fitting performance [1].

Disadvantage: need to choose hyperparameter mu_k [1] and s [1].

(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]

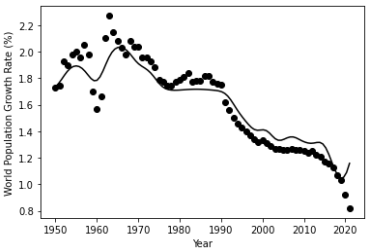


Figure 2 A

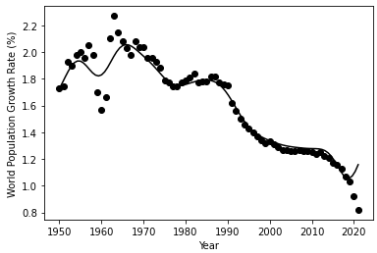


Figure 2 B

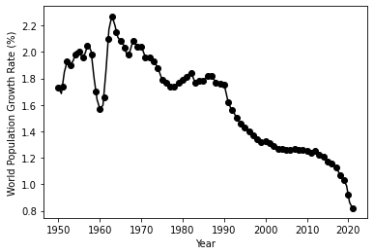


Figure 2 C

Figure 2A: Lasso [1], the fitted line misses many data points between 1990-2010, suggesting some weights of basis functions using these centres are pushed to zero [1]. Out of the three fitting methods, only Lasso with strong regularization can do this [1].

Figure 2B: Ridge regression [1]: the model ignores some extreme data points on the left [1], suggesting weights controlling the corresponding basis functions are very small [1].

Figure 2C: Linear regression [1]: the model fits most of the data points very well, especially fitting the data points between 1950 and 1970 perfectly [1], suggesting a large number of basis functions actively contribute to the fitted model [1].

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]**

1 mark for each feature and 1 mark for mentioning the feature type. Any reasonable feature is okay. For example, presence/absence of disrespectful words (binary feature), count of strongly negative words (real-valued scalar), word embeddings (real-valued vector).

(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]

Regular class accuracy: 100 %
Hate class accuracy: 0%
Weighted accuracy 50% (average across classes)

(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]

2 marks for calculating the linear combination
$z = w0 + w1x1+w2x2 = -1.8 + 2.1*1 -0.3*1 = 0$
2 marks for calculating the output probability score
$sigmoid(z) = 0.5$

(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]

This assumption lets LR learn only linear decision boundaries. Not suitable for data that requires a non-linear decision boundary - which is the case in many real problems. (1 mark for each observation)

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

Yes, possible. (1 mark for saying yes)
It resembles a sigmoid and shares some of the sigmoid's properties (bounded, monotonic). Not as smooth as sigmoid, but still will work. Optimization may be a bit more difficult (still possible through Linear Prog). (1 mark for commenting on resemblance with sigmoid and properties)

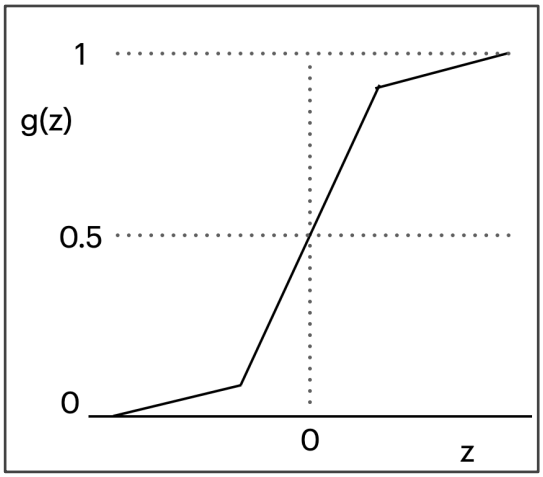


Figure 3: Proposed function

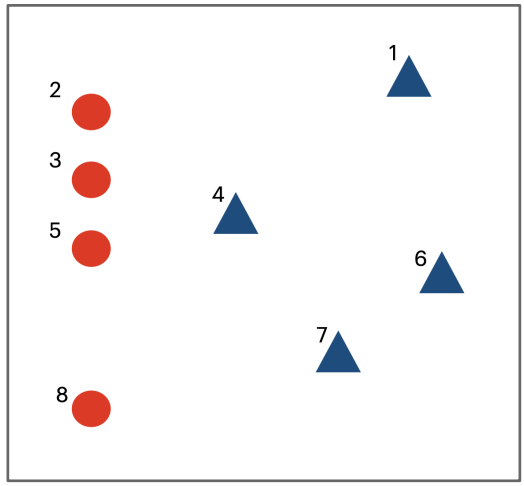
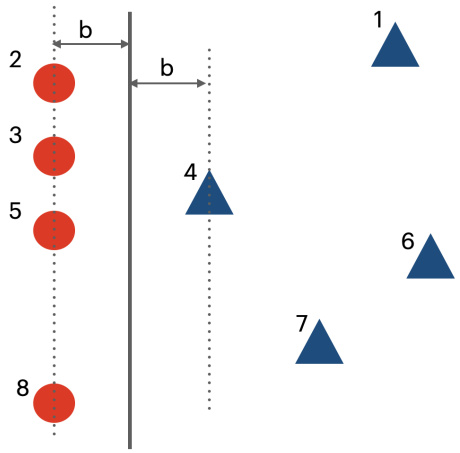


Figure 4: Training data points

(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]

Solution:



(ii) Which ones are the support vectors?
[2 marks]

2 marks for correctly finding all. 0 for missing even 1 support vector.
Support vectors: 2, 3, 4, 5, 8

(iii) What is the training error?
[1 mark]

1 mark for correct answer
Training error = 0

(iv) Removal of which data point will change the decision boundary?
[1 mark]

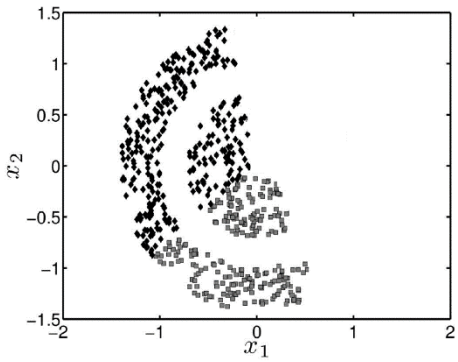
1 marks for correct answer
Data point 4.

3. Unsupervised learning question (Total marks 20)

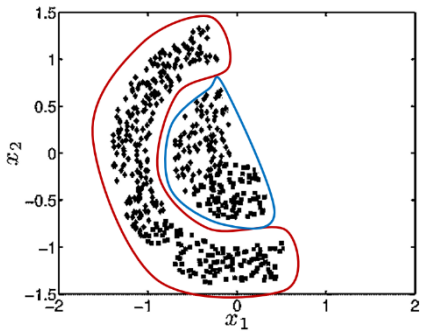
Consider using the K-means algorithm to perform clustering on the following scenario A1.

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]

K-means cannot split the data into two clusters (1 mark). Due to the euclidean distance points that are close together, although they belong to another manifold/cluster will be clustered together (1 mark).

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

[3 marks]

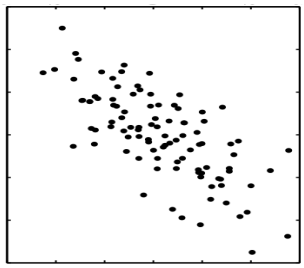
A kernel would help in this case (1 mark). A kernel would project the data onto a different space where data can be easily separated (1mark). A kernel also relaxes the dependency of k-means on Euclidean distance and it allow more appropriate distance measures to be used for the problem at hand (1 marks).

(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]

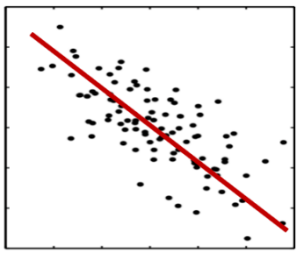
Mixture models might not be able to classify those data (1 mark). Mixture models assume a gaussian distribution of the underlying clusters and they capture different variations based on the parameters of the distribution (1 mark). In this case, it is not possible to approximate the complex boundary shape with a gaussian and the algorithm won't perform well (1 mark).

(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]

First principal component:

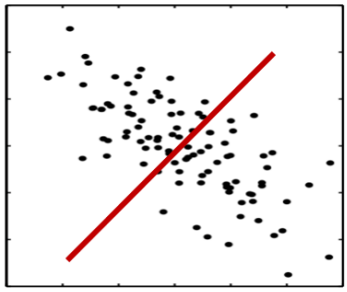


The first principal component will be across the direction of highest variance. (1 mark plot and 1 mark the explanation)

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

[2 marks]

PCA decompose the data into orthogonal components. Therefore, the second principal component will look: (1 mark: plot and 1 mark: explanation)



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

[2 marks]

PCA can be used for dimensionality reduction, intuitive visualization of high dimensional data and feature selection. (1 marks)

We can use cross-validation by leaving points at random and estimating the mean square error of the matrix factorization approach as we include more components. (1 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]

Projection methods will map data into different dimensions/coordinate system, and it will select the components based on the maximum variance. These methods are based on unsupervised learning, and they don't require class labels (1 mark). They project all features in a different space and therefore the new features are a combination of all the old ones (1 marks). Since they don't exclude completely specific features, they model better the intrinsic properties of the datasets in lower dimensions. (2 marks). They depend on the intrinsic properties of the data and their interrelationships. The disadvantage is that the number of optimum components might be unstable and less meaningful (2 mark).

Other points might be also acceptable: For example, if variables are highly correlated then feature selection techniques might not perform well.