University of Glasgow

Wednesday 08 December 2021
09.00 – 11.00 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 worth a total of 60 marks

1. Consider using regression to predict global temperature anomaly from cumulative CO2 emissions data showing in the following figure:

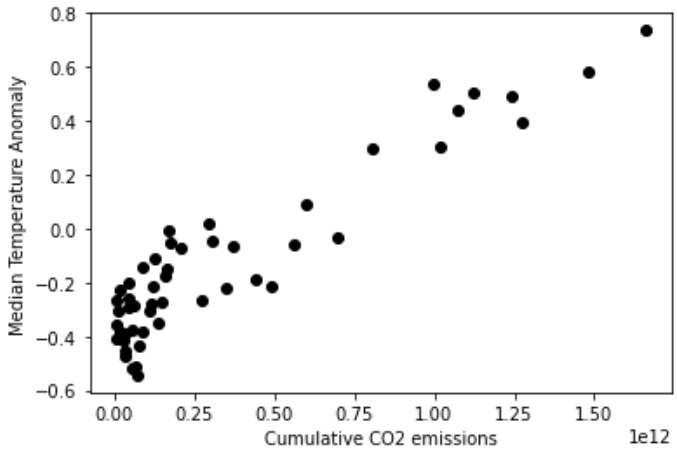

Figure 1. Global Temperature anomaly vs Cumulative CO2 emissions Data. Source: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

(a) Propose a rescaling strategy (with enough details of the procedure) for the cumulative CO2 emissions when using high order polynomial regression. Explain why the proposed strategy is appropriate.

[4 marks]

(b) Suppose a polynomial regression model with order of 1 is fitted to the data (without rescaling cumulative CO2 emissions). Identify a subset of data in figure 1 which will mostly likely be poorly fitted and explain why.

[6 marks]

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

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

where x_n represents each cumulative CO2 emission. Outline one advantage and one disadvantage of using RBF over polynomials for the data in figure 1.

[4 marks]

(d) Suppose we use the RBF in (c) with μ_k set to be the same as x_n, a commonly used approach in RBF, $s^2 = 1e24$, to fit the CO2/Temperature Anomaly 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]

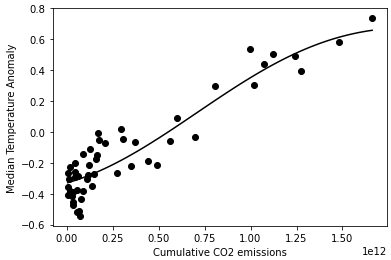

Figure 2 A

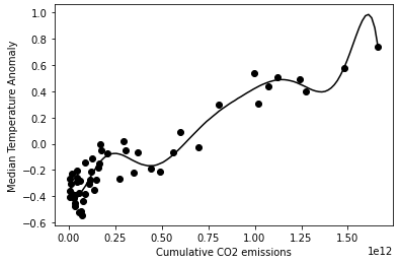

Figure 2 B

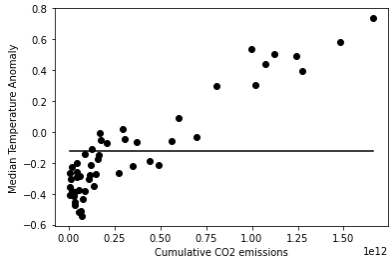

Figure 2 C

2. Classification question

(a) The likelihood of logistic regression is the following:

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

where $g(a) = \frac{1}{1+\exp(-a)}$. Consider the fitting this model to a dataset with 2 classes, 2 binary features and 2 examples per class:

Class 0: Example 1 = [1,1], Example 2 = [1,0].

Class 1: Example 1 = [1,1], Example 2 = [0,1].

Use the likelihood function to demonstrate which of the following two parameters hypotheses: [0.6, 0.1] and [0.6, 0.8] fits this dataset better.

[6 marks]

(b) Consider a support vector machine (SVM) is trained on a dataset where two data points are mislabeled by a non-expert annotator. The classifier outputs in the table below:

Correct label	0	0	0	1	1	1
Noisy label during training	0	1	0	1	0	1
Score of SVM	-9.6	8.8	0.7	?	2.2	0.3

(i) What would be the AUC (computed with the correct labels) if the missing value is 0.6? (Detailed calculation required)

[2 marks]

(ii) What would be the maximum achievable AUC (computed with the corrupted labels) and corresponding range of possible values for the missing value? Explain why.

[2 marks]

(iii) If you could correct one of the two corrupted labels to get better AUC (computed with the labels with one remaining wrongly labeled data), assuming the missing value is 0.6 and rest of the scores do not change. Which will you correct? Explain why.

[2 marks]

(c) Noisy labels may produce outliers in the training set. How will you configure the SVM in terms of margin and kernel to deal with outliers? Explain why?

[4 marks]

(d) Calculating AUC requires a classifier to give a score for each data point. A K-nearest neighbor classifier does not normally provide a score, but directly predicts the class for a data point. Outline two approaches to produce scores for computing AUC for a K-nearest neighbor classifier.

[4 marks]

3. Unsupervised learning question (Total marks 20)

Consider using the K-means algorithm to perform clustering on the following scenario in figure 3 A. We expect to form three clusters as shown in figure 3B.

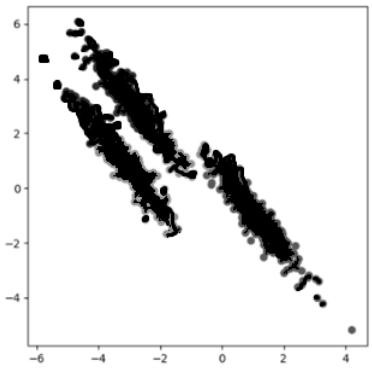

Figure 3 A Original Data

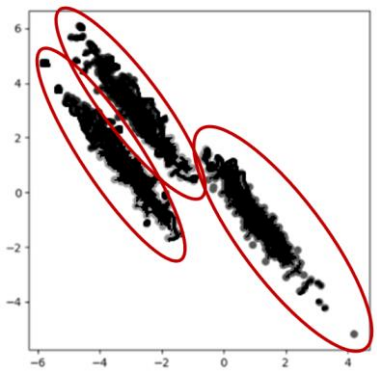

Figure 3 B: Expected Clusters

(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 help in this dataset and why?

[3 marks]

(c) An alternative approach is to use *mixture models.* Would mixture models help to better classify the dataset in figure 3 A than K-means and why?

[3 marks]

We want to cluster data in figure 4 A in three clusters as shown in figure 4 B.

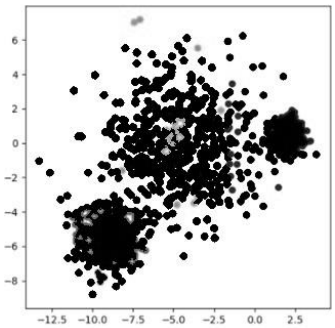

Figure 4 A: Original Data

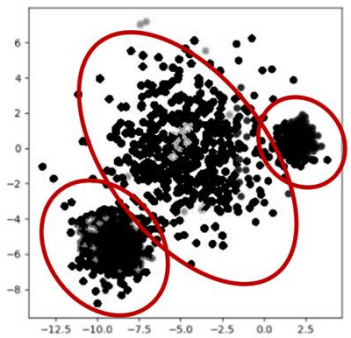

Figure 4 B: Expected Clusters

(d) 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?

[2 marks]

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

[3 marks]

(f) An alternative approach is to use *mixture models.* Explain whether mixture models could help to better classify this dataset and why?

[3 marks]

(g) Explain why there is a need for feature selection and list two methods and their main characteristics

[4 marks]