University of Glasgow

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

DEGREES OF MSc

Machine Learning & Artificial Intelligence for Data Scientists

(Answer all of the 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:

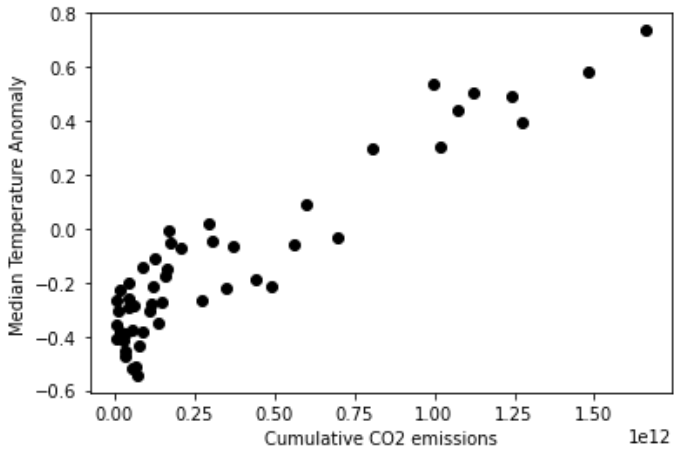

Figure 1. Global Temperature anomaly vs Cumulative CO2 emissions Data. Source: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

(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]

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 CO2 emissions, such that high order polynomial will still produce well behaved values (small) [1] and the matrix inversion in least square solution is still stable [1].

(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]

1 mark for identifying the correct poorly fitted data, which are the densely populated data points in the very left-hand side of the figure x valued in the range (0, 0.25e12). 5 marks for reasoning: polynomial regression model with order of 1 is a straight line [1], the data in figure could be fitted with two straight lines [1] one goes through the data in (0, 0.25e12) in x-axis [1], one goes through data from the very left to the very right of x-axis [1], the latter is likely to produce less average spare loss, leaving the data in (0, 0.25e12) in x-axis poorly fitted [1].

(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]

Advantage: the data is not equally distributed across x values, denser in small values and relatively sparser in large x values [1]. Using location specific basis functions RBF can model this localized effect better than polynomial functions which model global effect across all x values, resulting better fitting performance [1].

Disadvantage: RBF has more hyper-parameters [1], poorly chosen hyper-parameters could lead to overfitting [1].

(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]

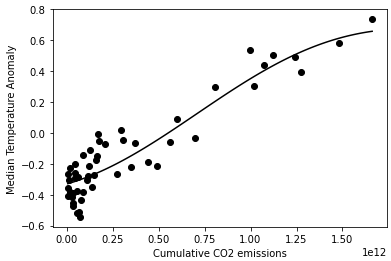

Figure 2 A

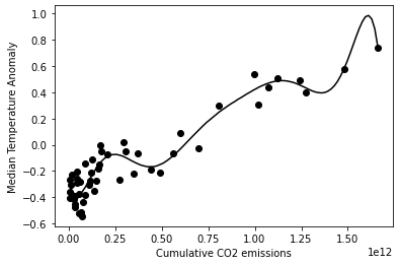

Figure 2 B

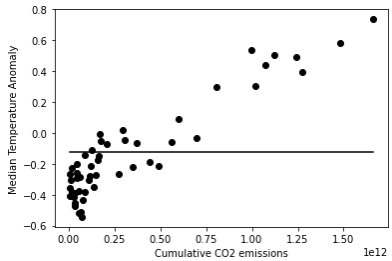

Figure 2 C

Figure 2A: Ridge regression [1]: the model ignores many densely populated data points on the left and points that could lead to bigger bends, suggesting weights controlling the corresponding basis functions are very small [1].

Figure 2B: Linear regression [1]: the model fits the densely populated data points on the left and the rest of the data very well, especially fits the two data points on the very right perfectly, suggesting large number of basis functions actively contribute the fitted model [1].

Figure 2C: Lasso [1], the fitted line is straight line parallel to the x-axis, suggesting all weights of the basis functions are zero. Out of the three fitting method, only lasso with very strong regularization can do this [1].

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]

Correct computation of each data likelihood [4 marks total]. Correct and consistent computation comparison of the total data likelihood for the two parameter hypotheses [2 marks].

Complete single data point and joint likelihood in log and normal scale:

Data index	Class	Feature 1	Feature 2	Log-likelihood of parameter candidate 1	Log-likelihood of parameter candidate 2	Likelihood of parameter candidate 1	Likelihood of parameter candidate 2
Data point 1	0	1	1	-1.103186	-1.604174	0.33181223	0.19781611
Data point 2	1	1	1	-0.403186	-0.220474	0.66818777	0.80218389
Data point 3	0	1	0	-1.037488	-1.037488	0.35434369	0.35434369
Data point 4	1	0	1	-0.6443967	-0.3711007	0.52497919	0.68997448
			Joint log-likelihood:	-3.1882567	-3.2494235		
					Joint likelihood:	0.04124370801	0.03879656939

Parameter candidate [0.6, 0.1] fits the data better.

(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]

4/9 [1] (1+2+1)/(3*3) [1]

(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]

AUC 7/9, > 2.2. These numbers ensure that positive data (based on corrupted labels) have high score than any negative data (based on corrupted labels label).

(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]

The one with score of 2.2. The other corrupted label with score of 8.8 with result in much lower AUC.

(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]

Soft margin, to allow outliers to go across the decision boundary [2]. Kernel, choose a less powerful kernel to avoid overfitting [2].

(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]

2 marks each, for example, converting vote counts to vote proportions and using majority margin.

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.

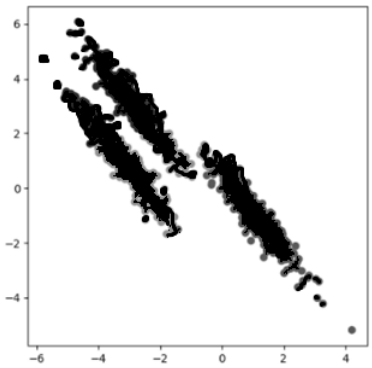

Figure 3 A Original Data

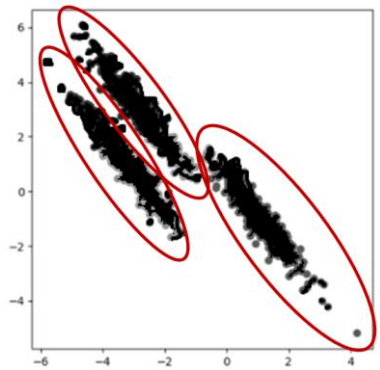

Figure 3 B: Expected Clusters

(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]

K-means cannot split the data into three clusters along the respective ellipsoids. Due to the euclidean distance points that are close together but in neighbouring ellipsoids will be clustered together. (One mark for the answer and one for the explanation)

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

[3 marks]

A kernel projects the data onto a different space where data can be easily separated. The space could have a higher or lower number of dimensions compared to the original data. (One mark that this is possible and two marks for the explanation)

(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]

Mixture models should be able to better classify the data than k-means, since there are able to model clusters as a mixture of gaussian distributions with anisotropic gaussian distribution (diagonal elements of covariance matrix are not equal)

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

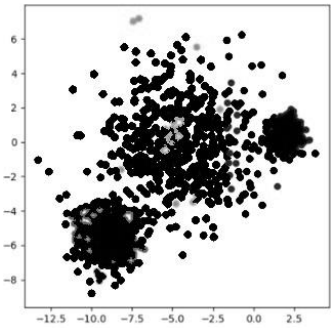

Figure 4 A: Original Data

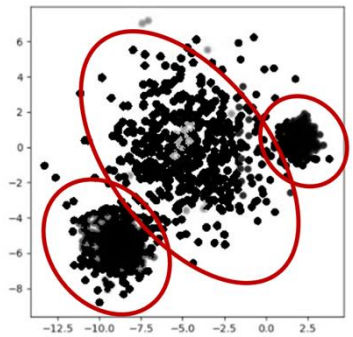

Figure 4 B: Expected Clusters

(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]

K-means cannot split the data well into three clusters because the variance in the middle cluster is considerably different than the variance in the other clusters. Therefore, considering only distance won't be sufficient.

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

[3 marks]

Kernel K-mean does not explicitly model variance and since it is based on distance it won't be robust in classifying data with anisotropic variance.

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

[3 marks]

Mixture models with anisotropic variance would work well to model these data since variance in the data is a parameter for each cluster.

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

[4 marks]

Due to the curse of dimensionality, which states that the number of required samples increases exponentially with the number of features, it is desirable to reduce dimensionality. Also it is important for visualising data and identifying anomalies.

One strategy is to use a subset of the originals (ie.choose those features that maximise the difference between the two classes)

Another strategy is to combine the original and find new dimensions (ie. dimensions that maximise the variance -- PCA)

(Explanation two marks and one mark for each strategy)