Exam Feedback COMPSCI 5100 Machine Learning & Artificial Intelligence for Data Scientists 2021/2022

Q1 (a)

General comments:

This is a question about feature rescaling for polynomial regression. It has been discussed in lectures and practiced in labs of the linear regression unit.

Most answers can propose a reasonable strategy. Many fall short in connecting why the strategy is suitable for the CO2 emission vs Temperature data.

Model answer:

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].

Q1 (b)

General comments:

This is a question about the effect of outliers on fitting a basic straight line.

There are two diverging trends in the figure: 1) data points concentrated on the left; 2) data points spread over the x-axis. Most of answers can identify one trend, but often with inadequate discussion.

Model answer:

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]

Q1 (c)

General comments:

This is a question about the difference between RBF and polynomial regression. The difference and their consequence were discussed in linear regression lectures. Here, we expect the answer to be put in the context of the CO2 data.

Most answers can list the features of RBF generally. Very few attempted any meaningful comparison with polynomial regression. Notably, many don't view RBF and polynomial

regression as linear regression. They are both linear regression models, just with different basis functions.

Model answer:

Advantage: the CO2 data is not equally distributed across x values, denser in small values and 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].

Q1 (d)

General comments:

This is a question about regularisation in linear regression. This topic was discussed in detail in linear regression unit.

The most notable error here is that many argue the straight line fit in figure 2c as linear regression and others as not.

Model answer:

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].

Q2 (a)

General comments:

This is an application of the likelihood function taught in logistic regression. The question provided the formula of the likelihood function, and values for each variable in the formula. We expect the answers to plug the values in and use the results to compare the results.

Most answers can compute x*w, but very few can evaluate the rest of the likelihood function.

Model answer:

There are two ways to make the comparison: likelihood function or log likelihood function. We list the complete solution in both. 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	-2.671644692	-4.018149928	0.06913842034	0.01798620996
Data point 2	0	1	0	-2.671644692	-4.018149928	0.06913842034	0.01798620996
Data point 3	1	1	1	-0.07164469197	-0.01814992792	0.9308615797	0.98201379
Data point 4	1	0	1	-0.07164469197	-0.01814992792	0.9308615797	0.98201379
			Joint log-likelihood:	-5.486578768	-8.072599712		
					Joint likelihood:	0.004141990673	0.0003119711908

Parameter candidate 1 [0.6, 0.1] fits the data better.

Q2 (b)

General comments:

Three questions about applying the knowledge of computing AUC with label and scores. The topic is covered in lectures in classification unit. Two questions similar in style from past and mock exam papers were discussed in the revision lecture.

Model answer:

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

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

Q2 (c)

General comments:

This is a direct application of soft margin and kernel in SVM, which was discussed in the classification unit. This question is generally well answered.

Model answer:

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

Q2 (d)

General comments:

The very discussion that can provide the perfect answer happened when KNN was introduced in classification lectures.

Most answers can lay out one way of computing a score for AUC.

Model answer:

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

Q3

General Comments:

Question 3 explores the limitations of clustering techniques with relation to the distribution of the data.

Q3(a,b,c) and Q3(d,e,f) have been discussed in the lecture with relation to the limitations of k-means clustering. The advantages of mixture models with relation of modeling gaussian distributions have also been presented. The question tested critical thinking and understanding of the high-level function of the clustering techniques presented in the class. This is necessary to pick the appropriate method for the data at hand.

Finally, Q3-g asked for the reasons behind feature selection and two methods along with their main characteristics. This information was covered in the lecture about projection. Furthermore, it was related to most of the coursework tasks.

This question was generally answered well. The students have understood well the limitations of k-means and the advantages of mixture models. It was more difficult to reason behind kernel k-means advantages and limitations.

Model Answer:

A) 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. [2 marks]

B) A kernel that 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. [3 marks]

C) 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). [3 marks]

D) 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. [2 marks]

E) Although the kernel K-mean approach is able to project the data onto a different space where data can be easily separated, considering only distance it is not sufficient to model the difference in the variance in the data. [3 marks]

F) Mixture models with anisotropic variance would work well to model these data since variance in the data is a parameter for each cluster. [3 marks]

G) 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. [2 marks]

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) (Other possible answers exist and they could take full marks as soon as they are backed with appropriate arguments) [2 marks]