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:

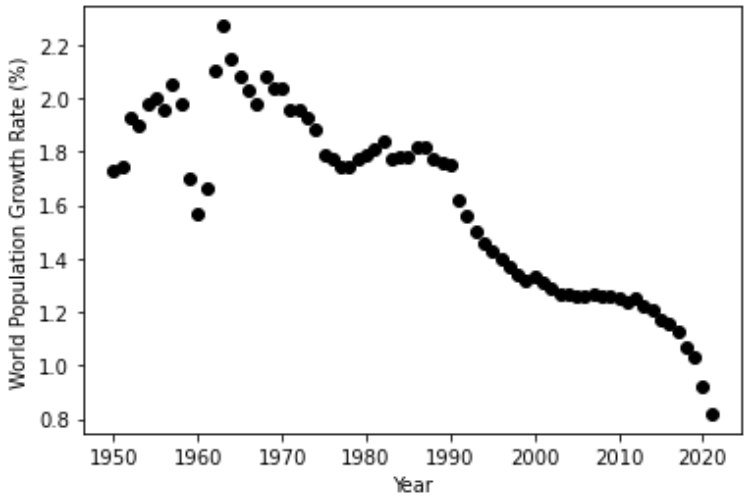


Figure 1. World population growth rate from 1950 to 2021. Source: https://ourworldindata.org/world-population-update-2022

(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]

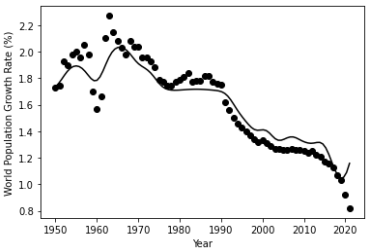


Figure 2 A

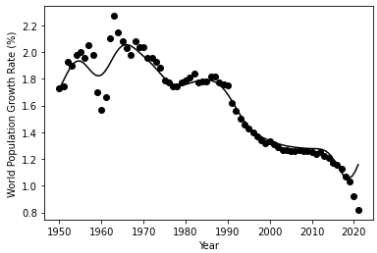


Figure 2 B

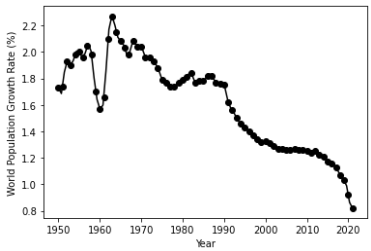


Figure 2 C

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)

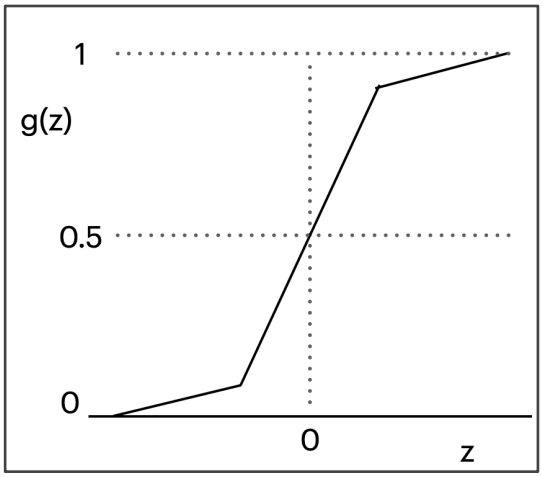


Figure 3: Proposed function

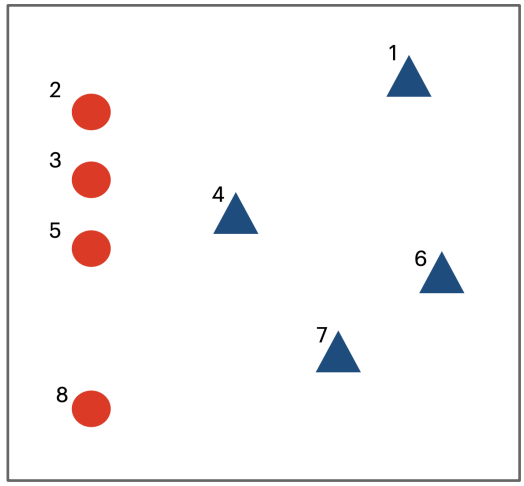
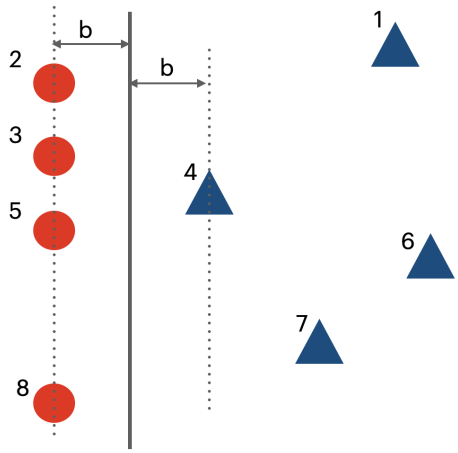


Figure 4: Training data points

(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]

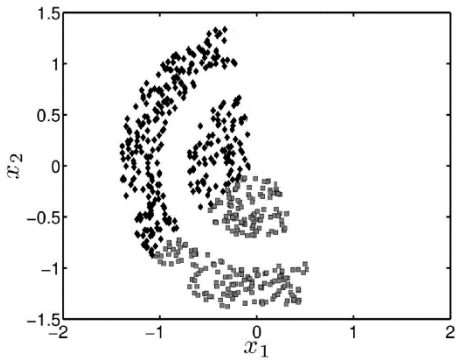
1 marks for correct answer
Data point 4.

3. Unsupervised learning question (Total marks 20)

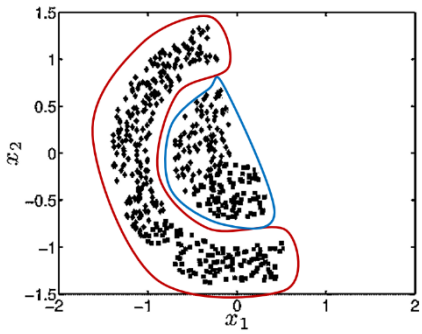
Consider using the K-means algorithm to perform clustering on the following scenario A1.

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]

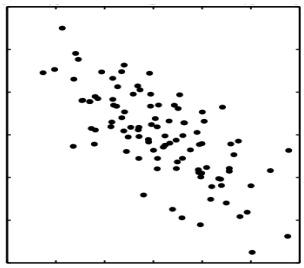
A kernel would help in this case (1 mark). A kernel would project the data onto a different space where data can be easily separated (1mark). A kernel also relaxes the dependency of k-means on Euclidean distance and it allow more appropriate distance measures to be used for the problem at hand (1 marks).

(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]

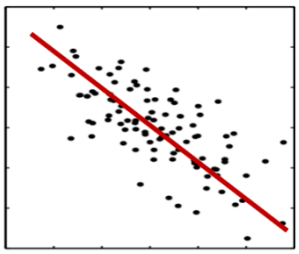
Mixture models might not be able to classify those data (1 mark). Mixture models assume a gaussian distribution of the underlying clusters and they capture different variations based on the parameters of the distribution (1 mark). In this case, it is not possible to approximate the complex boundary shape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

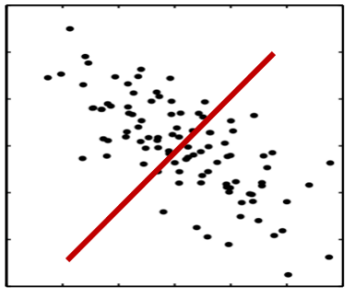


The first principal component will be across the direction of highest variance. (1 mark plot and 1 mark the explanation)

(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.

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:

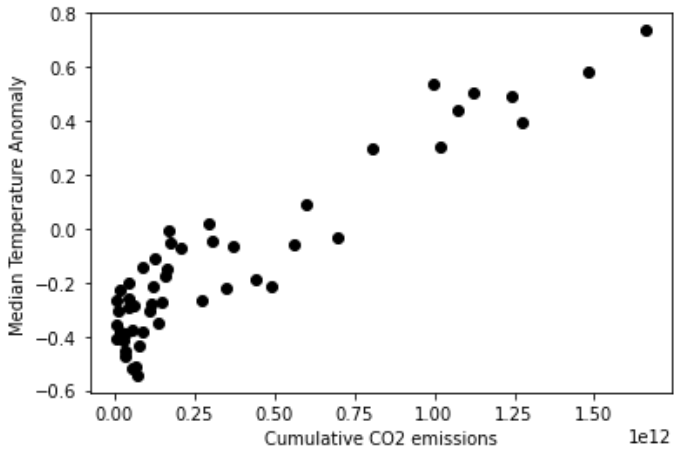


Figure 1. Global Temperature anomaly vs Cumulative CO2 emissions Data. Source: https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions

(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]

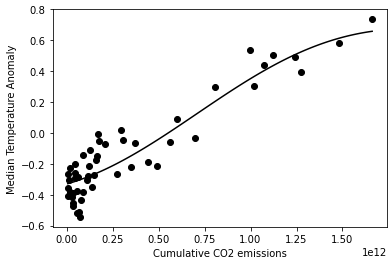


Figure 2 A

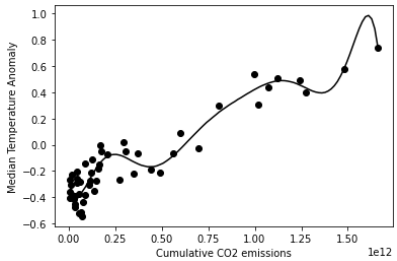


Figure 2 B

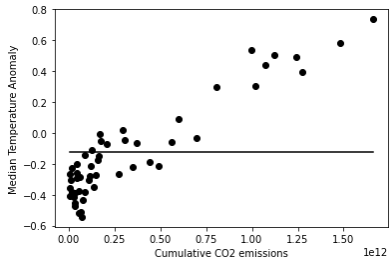


Figure 2 C

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.

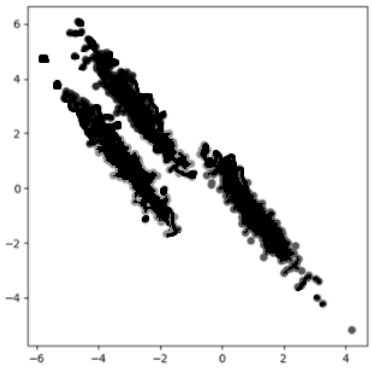


Figure 3 A Original Data

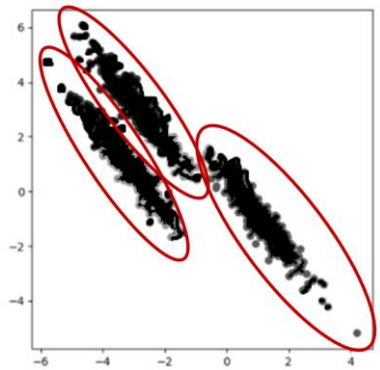


Figure 3 B: Expected Clusters

(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.

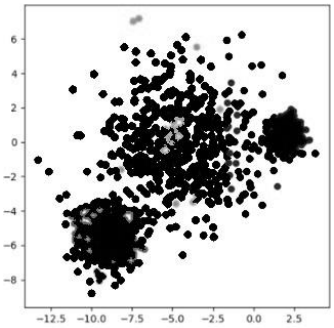


Figure 4 A: Original Data

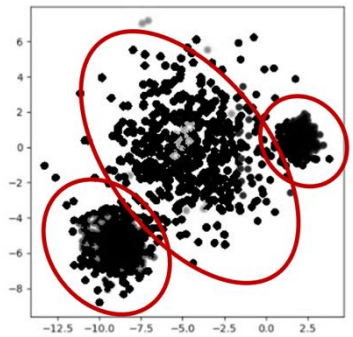


Figure 4 B: Expected Clusters

(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)

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.

1. Considering linear regression on the Olympic data in figure 1.

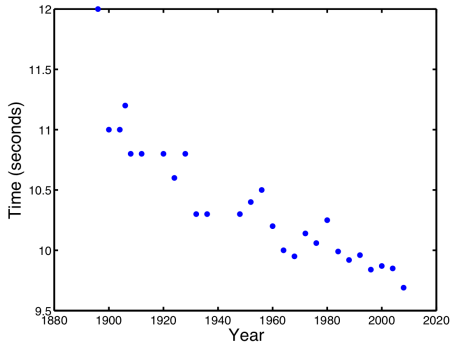


Figure 1: Olympic data

(a) We want to predict Olympic years from 100m winning times. What should be the target value and attribute? When solving this regression task with a polynomial regression model, how would you rescale the attributes? Why?

[6 marks]

Target: years [1], attribute: winning times [1]. A reasonable solution with sufficient details. E.g. whiting (x-mean(x))/std(x) [2]. A reasonable explanation of what the solution can do. E.g. Whiting makes sure the attribute is in [2].

(b) Based on what you have learned from fitting linear regression models (with polynomial or RBF) to the relationship between years and winning times, predict which year may produce winning time 9s and 13s, explain why.

[4 marks]

Answer should include reasonable estimation of years that may produce winning time 9s and 12s [2], using arguments from existing data and model, could be polynomial or RBF [2].

For example, polynomial order of 3 might be a good fit to the data, the model is likely to predict 9s after 2040, 13s could be before 1860.

(c) The radial basis function (RBF):

$$h_{n,k} = \exp\left(-\frac{\sum_{d=1}^{D}(x_{n,d} - \mu_{d,k})^2}{2s^2}\right), n = 1, \dots, N;\ k = 1, \dots, K$$

is a popular basis function. The parameter $\mu_{d,k}$ is often be a data point $x_{i,d}, i = 1, \dots, N$. Outline the strength and risk of this setup for $\mu_{d,k}$, and how would you mitigate the risk.

[5 marks]

Strength: flexibility [1]. Risk: numerical stability and overfitting [2]. Use less centers or add small value to the diagonal of X^TX, using regularization [2].

(d) In addition to the polynomial function and RBF, linear regression can be generalized using other basis functions. One of most widely used example is the Fourier analysis, let's consider the following linear regression model:

$$t_n = \sum_j^m A_j \cos(jx_n + \theta_j)$$

What is the basis function of choice here? How would you deal with the unknow parameters A_j and θ_j? (Hint: you might find the following trigonometry identity useful, $\cos(a+b) = \cos(a)\cos(b) + \sin(a)\sin(b)$).

[5 marks]

Solution 1[3 marks in total]: Basis function: for $\cos(jx_n + \theta_j)$[1], A_j is the regression parameter, cross validation for \theta [2]

Solution 2 [full mark]: Two basis functions as a result of applying the provided identity $\cos(jx_n)$ and $\sin(jx_n)$. \theta_j becomes part of the linear regression parameter, the same as A_j.

2. Classification question

(a) The likelihood of logistic regression

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

where $g(a) = \frac{1}{1+\exp(-a)}$. Use an example of a few data points to explain how the likelihood function tells how well the parameter $\mathbf{w}$ fits the data.

[4 marks]

The example needs to have more than one pair of t_n and x_n [1], demonstrating how to construct joint data likelihood [1]. It should also include a parameter estimate representing a good fit and a parameter estimate representing a bad fit [2].

(b) The following matrix contains estimated parameters values from three types of logistic regression models. The model type is indicated by the columns. The parameter of each feature is placed in the corresponding row. Give your best estimate of what each model is and explain why.

Model 1	Model 2	Model 3
[[1.08381535e+01	1.19648635e+01	1.11285803e+01]
[-0.00000000e+00	-1.29443055e+01	-3.29359603e-01]
[-0.00000000e+00	5.79522897e+01	-1.94725736e-01]
[-1.16126582e-01	-1.09582035e+02	-9.64898104e-02]
[-1.59001968e-02	6.23248849e+01	-1.87327081e-02]
[-0.00000000e+00	7.48519704e+01	3.32402164e-02]
[1.38119952e-03	-1.46955431e+02	4.50182751e-02]
[3.22128802e-03	1.04735797e+02	9.53751777e-03]
[1.61616847e-04	-3.88035781e+01	-3.60588365e-02]
[-8.65262203e-05	7.43343695e+00	1.40369595e-02]
[-7.74413350e-05	-5.82870289e-01	-1.62830483e-03]]

[6 marks]

Model 2 [2]: logistic polynomial regression. Some parameter values are very big in absolute value. Model 3 [2]: L2-regularised logistic regression. Compare to model 1, most parameters are much smaller in absolute value. Model 1 [2]: L1-regularised logistic regression. Some parameters are exactly zeros.

(c) Compare the effect on prediction of the three logistic models in **(b)**.

[4 marks]

With the same x_n [1], L1- and L2-regularised logistic regression are likely to produce lower probability of being the positive class [2], they have better generality with unseen data [1].

(d) Let's consider a binary classifier trained on a falsely labeled dataset. The issue is all positive (*1*) and negative (*0*) labels are swapped during training. The classifier outputs in the table below:

Correct label	0	0	0	1	1	1
False label during training	1	1	1	0	0	0
Probability of being the *positive* class	0.9	0.8	0.7	?	0.2	0.1

(i) What would be the AUC (computed with the correct labels) when the classifier is perfectly trained on the false data? And why?

[2 marks]

0, Perfect AUC on false label is 1. Correct label is 1-1. Or arguing using the definition of AUC: the probability of a positive example having classification score higher than a negative one.

(ii) Provide the range of possible values for the missing output (labeled '?') that would be produced by the classifier in (i). Explain why.

[2 marks]

[0, 0.7), all number ensures that all positive data (based on false label) have high score than any negative data (based on false label).

(iii) What would be the AUC (computed with the correct labels) of a random classifier trained on the falsely labeled data? Why?

[2 marks]

0.5, random classifier will have 0.5 AUC on false label, 1-0.5 is still 0.5.

3. Clustering question (Figures in this question were taken from the sklearn clustering tutorial: https://scikit-learn.org/stable/modules/clustering.html)

(a) Describe clustering results of K-means and Gaussian Mixture in figure 2. Hint: answer should address parameters estimation, initial conditions and selecting the number of clusters.

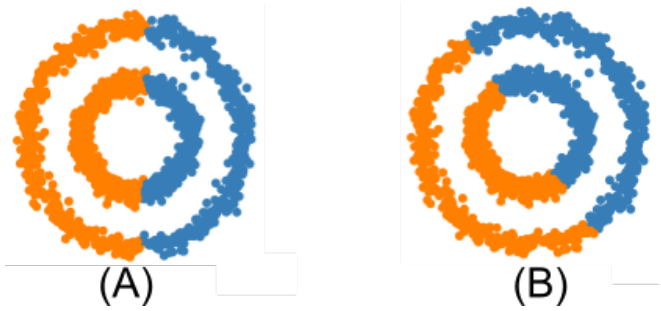


Figure 2: Clustering results of (A) K-means and (B) Gaussian mixture model.

[6 marks]

3 marks each: answer should address parameters estimation [1], initial conditions [1] and selecting the number of clusters [1].

Example solution:

Kmeans: two cluster centers are located somewhere between the two rings (closer to the inner ring) [1]. It's hard to determine what initial conditional might have been used. Many splits are possible. [1] Difficult to do model selection for the same reason (Many splits are equally well) [1].

GMM: the means are also located somewhere between the two rings. The two covariance matrices and mixing weights should be almost identical between the two [1]. Similar to Kmeans, there are many equally good solutions for GMM. Specific initial condition is hard to determine [1]. Similar to Kmeans, model selection is hard due to multiple good solutions in each number of cluster setting [1].

(b) Suppose we want to avoid any data point from the inner ring being assigned to the same cluster with any point data point from the outer ring. Outline two approaches to achieve this goal with the Gaussian mixture model? Hint: You don't have to use just 2 clusters.

[4 marks]

Project the data onto to a different space [1] where data points in the inner ring are well separated from the data points in the outer ring [1]. Restrict variance of each component to be small [1] and use more clusters [1].

Marks for any other sensible approaches

(c) Describe clustering results of K-means and Gaussian Mixture in figure 3. Hint: answer should address parameters estimation, initial conditions and selecting the number of clusters.

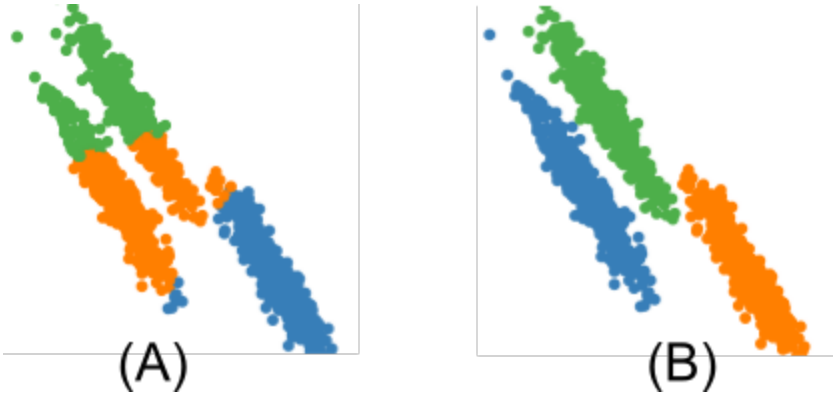


Figure 3: Clustering results of (A) K-means and (B) Gaussian mixture model.

[6 marks]

3 marks each: answer should address parameters estimation [1], initial conditions [1] and selecting the number of clusters [1].

Example:

Kmeans: three cluster centers are vertically distributed [1]. Initial cluster centers may also have been vertically distributed while others are possible [1] A cross validation on clustering number might prefer the number of clusters larger than 3. [1].

GMM: three highly Gaussian-like distributed data. Different means but similar covariance matrices and mixing weights [1]. In this case, the optimal solution is quite clear, therefore, GMM should be same with multiple initial with only the identity the clusters changing [1]. Cross-validation or Marginal likelihood should favor a 3 cluster solution [1].

(d) Suppose Figure 3 (B) represents the results we want. Outline one approach to achieve this goal with K-means. Hint: Sufficient details of the approach are required to get full marks.

[4 marks]

Kernel k-means [2]. Details on how to kernelize K-means: using the kernel trick on the distance [1] and only update the assignments [1].

University of Glasgow

DayOfWeek DayOfMonth Month 2XXX
XX.XX am/pm – XX.XX am/pm
(Duration: 60 minutes)

DEGREES OF MSc, MSci, MEng, BEng, BSc,MA and MA (Social Sciences)

Machine Learning 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.

1. A polynomial regression model is defined as:

$$t_n = \sum_{d=0}^{D} w_d \, x_n^d, n = 1, \ldots, N$$

(a) When applying this model to the Olympic data, where $x_n \in 1896, \ldots, 2008$, we always rescale x, e.g. $x_n = \frac{x_n - 1896}{40}$. Explain why is this rescaling necessary.

[4 marks]

In the Olympic data, high polynomial could make the value of x^d very big [2]. As a result, the computation of parameters becomes unstable [2].

(b) Write down the regression model if the following Radial basis function (RBF) with a range of different position parameter μ is applied to x_n.

$$RBF(y; \mu, l) = \exp(-\frac{(y-\mu)^2}{l^2})$$

[3 marks]

$$t_n = \sum_{d=1}^{D} w_d \exp(-\frac{(x_n - \mu_d)^2}{l^2}), n = 1, \ldots, N$$

(c) Give an example of nonlinear regression model. Also explain why it is nonlinear.

[2 marks]

Any nonlinear function of the w and x e.g. f(x, w) = sin(w*x) [1]. State clearly which relationship is nonlinear [1].

(d) Explain why polynomial regression could suffer from outliers.

[3 marks]

Outliers favours higher polynomial orders [1], as they can fit the outliers better [1]. However, the high order models are more likely to overfit the data [1].

(e) L2 Regularised regression can be used to deal with this problem. Use a contour plot (Assuming the dimension of the parameter is 2) of parameters and the loss function to explain why.

[8 marks]

2 marks for the correct contour of the mean squared error, 2 marks for the correct contour for the L2 regularisation. 2 marks for highlight the correct intersection between the two. 2 marks for stating the fact that optimal parameter shifted closer to (0, 0)

2. Classification question

(a) Classification and regression are both supervised learning problems. Describe a way to turn a regression problem into a classification problem? (Please state the differences between the two.)

[3 marks]

In regression, the target variable is a continuous/real-valued variable [1]. In classification, the target variable is a discrete/binary/categorical variable [1]. To obtain a classification problem from regression, one can cluster/group the continuous/real-valued variable into groups [1].

(b) The receiver operating characteristic (ROC) curve is a standard way to visualize the performance of classifiers. Outline how to draw a ROC curve

[3 marks]

The ROC curve is created by varying the threshold at which the classifier calls something as belonging to the positive class. [1] Each point is consisted of false positive rate or 1-specifity (normally on the x-axis) [1] and true positive rate or sensitivity (normally on the y-axis) [1].

(c) For the classifier outputs in the table below, provide a value for the missing output (labeled '?') that would:

Class Label	0	0	0	1	1	1
Output	0.1	0.25	0.4	?	0.6	0.9

(i) Give an AUC equal to 1,

[2 marks]

Anything between 0.4 and 1.0

(ii) Give an AUC less than 1,

[2 marks]

Anything less than 0.4

(iii) Now assuming you can change any output of the six data points, give an example of the outputs, such that the AUC is 0.5.

[3 marks]

As long as the output for 0s and 1s are the same.

(d) Use a diagram (some data in 2D) to describe how linear Support Vector Machines (SVMs) operate (how they make classification decisions, what and how parameters have to be set, what data needs to be stored etc).

[4 marks]

1 mark for drawing that demonstrate the margin, 1 mark for highlight the decision boundary, 1 mark for highlighting support vectors, 1 mark for pointing out only support vectors needs to be stored.

(e) Logistic regression uses the sigmoid function to make classification decision. Now, we have defined the following probability with a sigmoid function.

$$p(t_n = 0|\mathbf{w}, \mathbf{x}_n) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x}_n)}$$

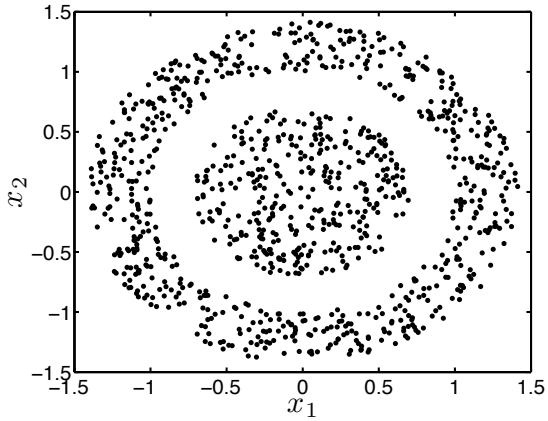
Write down the corresponding likelihood function for nth data points, $p(t_n|\mathbf{w}, \mathbf{x}_n)$.

[3 marks]

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

3. Unsupervised learning question

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



We want to cluster data in the outer ring in one cluster and the data in the inner circle as a different cluster.

(a) Outline what would happen if we directly apply *K*-means with Euclidian 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 into outer ring and inner circle clusters [1]. It will group parts of the outer ring or inner circle [1].

(b) An alternative approach is to use *Kernel K-means*. Explain how the kernel could help in this dataset.

[3 marks]

A kernel that project the data onto a different space [1] where data can be easily separated [1]. The space could have higher or lower number of dimensions compare to the original data [1].

(c) Which one of the following statements about kernel is NOT correct?

A. One could use the RBF kernel, $K(\boldsymbol{x}_n, \boldsymbol{x}_i) = \exp(-\gamma(\boldsymbol{x}_n - \boldsymbol{x}_i)^T(\boldsymbol{x}_n - \boldsymbol{x}_i))$, to achieve the clustering target.

B. The RBF kernel projects the data onto an infinite dimensional space. The free parameter γ can be estimated with cross-validation.

C. One could use the linear kernel, $K(\boldsymbol{x}_n, \boldsymbol{x}_i) = \boldsymbol{x}_n^T \boldsymbol{x}_i$, to achieve the clustering target.

D. The linear kernel projects the data onto itself, and there is no free parameter to tune.

[2 marks]

C

(d) Write some pseudo-code to perform K-means.

[5 marks]

Given: Number of clusters, K

2. For each cluster k = 1...K:

3. For each object n = 1...N: [1]

4. Compute the distance between object n and cluster k [1]

5. Assign object n to the cluster corresponding to the smallest distance [1]

6. Update the mean of each cluster [1]

7. If assignments have changed, return to 2. Else stop. [1]

(e) Outline how and why cross-validation can be used to select the number of clusters K.

[8 marks]

Cross-validation with total or average Euclidean distance between data points and their cluster centers [2]. At each CV cycle, use the training data to determine the mean of the clusters [2]. Test these means by computing total or average Euclidean distance between testing data points and their nearest cluster centers [2]. CV allow the trained number of clusters and means to be tested on a different dataset in which the trained K-mean may or may not be a good fit. [2]

MLAI4DS Mock Exam Paper

This examination paper is worth a total of 60 marks

1. Linear regression with models of the form $t_n = \mathbf{w}^T\mathbf{x}_n$, is a common technique for learning real-valued functions from data.

(a) Squared and absolute loss are defined as follows:

$$L_{squared} = (t_n - \mathbf{w}^T\mathbf{x}_n)^2, \quad L_{absolute} = |t_n - \mathbf{w}^T\mathbf{x}_n|$$

Describe, with a diagram if you like, why, when optimizing the parameters with the squared loss outliers have a larger effect than with the absolute loss.

[6 marks]

Defining e_n as t_n –w^Tx_n, for values of |e_n| < 1, the absolute loss is larger than the squared loss [2]. However, as e_n increases, the squared loss increases much faster [2]. Outliers have a high e_n and therefore a much larger influence over squared loss than in absolute loss[2]

(b) Which of the following statements is true:

A) Parameter estimation with the squared loss is not analytically tractable.

B) The squared loss is equivalent to assuming normally distributed noise.

C) The absolute loss is a popular choice for regularization.

D) The squared loss is a popular choice for regularization.

[2 marks]

B

(c) Discuss why the value of the squared loss on the training data cannot be used to choose the model complexity.

[3 marks]

When training the model we are minimizing the loss on the training data [1]. As we make models more complex, the squared loss value at the minima will always decrease [2]. Hence, training loss will always favour more complex models.

(d) For the particular model $t_n = w_1 x_n + w_2 x_n^3$, I optimize the parameters and end up with $\mathbf{w} = [2,1]^T$. What does the model predict for a test point at $x_{new} = 3$?

[2 marks]

2*3 + 3^3 = 33

(e) The radial basis function (RBF):

$$h_{n,k} = \exp\left(-\frac{\sum_{d=1}^{D}(x_{n,d} - \mu_{d,k})^2}{2s^2}\right), n = 1, \ldots, N;\ k = 1, \ldots, K$$

is a popular choice for converting the original features, $x_{n,d}$, into a new set of K features

prior to training. Assume the value of s is given. Describe a procedure for determining the center parameter $\mu_{d,k}$ and K.

[4 marks]

Any reasonable approach for choosing K centers, the answer should clearly state how to choose K [2] and what is $\mu_{d,k}$[2]. A common choice is let $\mu_{d,k}$ be a data point $x_{i,d}, i \neq n$ other than the nth data point. In this case, K = N-1.

(f) With respect to the functions they can fit, describe the difference between RBF and the basic linear model $\mathbf{w}^T\mathbf{x}_n$ with a graph.

[3 marks]

The RBF can produce a nonlinear model i.e. wiggly curve [1]. $\mathbf{w}^T\mathbf{x}_n$ produces only the linear model i.e. straight lines or hyperplanes [1]. 1 mark for readable graph.

2. Classification question

(a) Use a classification algorithm, describe what is meant by:

(i) Generalisation

[2 marks]

(ii) Over-fitting

[2 marks]

Plenty of sensible answers. Marks awarded for a description of generalization that describes the model's ability to make predictions on previously unseen data and, for overfitting, the problem of *memorizing* the training data.

(b) A classification algorithm has been used to make predictions on a test set, resulting in the following confusion matrix:

		Truth		
Prediction		**Positive**	**Negative**	**Total**
	Positive	23	5	28
	Negative	10	12	22
	Total	33	17	50

Compute the following quantities (expressing them as fractions is fine):

(i) Accuracy

[2 mark]

35/50

(ii) Sensitivity

[2 mark]

23/33

(iii) Specificity

[2 mark]

12/17

(c) Explain why it is not possible to compute the AUC from a confusion matrix.

[4 marks]

AUC is the area under the ROC curve [1]. The ROC curve is created by varying the threshold at which the algorithm calls something as belonging to the positive class [1]. A confusion matrix only gives us the performance at one threshold [1].

(d) Two binary classifiers are used to make predictions for the same set of six test points. These predictions are given below, along with the true labels. Compute its area under the curve (AUC) in each case.

Classifier 1		Classifier 2	
Predicted probability of class 1 (Score of class 1)	True class	Predicted probability of class 1 (Score of class 1)	True class
1	1	0.8	1
0.8	1	0.8	0
0.6	1	0.6	1
0.4	0	0.6	0
0.2	0	0.2	1
0.0	0	0.2	0

[4 marks]

AUC for classifier 1 is 1 [2] and 0.5 for classifier 2 [2].

(e) Explain how the SVM can be extended via the kernel trick to perform non-linear classification.

[2 marks]

In the SVM objective function, the data only appear in the form of inner products [1]. Inner products in the original space can be replaced by inner products in another feature space via kernel functions [1].

3. Unsupervised learning

(a) Provide pseudo code for K-means (assume that the number of clusters is provided).

[5 marks]

Given: Number of clusters, K

2. For each cluster k = 1...K:

3. For each object n = 1...N: [1]

4. Compute the distance between object n and cluster k [1]

5. Assign object n to the cluster corresponding to the smallest distance [1]

6. Update the mean of each cluster [1]

7. If assignments have changed, return to 2. Else stop. [1]

(b) Is the total Euclidean distance between data points and their cluster centers a good criterion to select number of clusters in K-means? Why?

[2 marks]

No [1], large number of clusters will lead to smaller and better Euclidean distance [1]

(c) Gaussian mixture models can be fitted to data using the expectation maximization (EM) algorithm. The EM algorithm has two steps: E-step and M-step. Describe what parameters are being estimated in each step in Gaussian mixture models.

[4 marks]

In the Estep, the estimated parameters are the expected assignment probabilities for each data point to each Gaussian component [2]. In the Mstep, the mean and covariance of Gaussian components and the mixing coefficients [2].

(e) Describe three key differences between K-means and Gaussian mixture models

[4 marks]

K-means use hard assignment [1]. Gaussian mixture models estimate the probability of assignment i.e. soft assignment [1]. GMM estimate both cluster center [2] and covariance [1].

(e) K-means often converges to a local optimal solution. Describe a simple process for overcoming the local optimality of K-means.

[3 marks]

For fixed K, perform multiple re-starts, and evaluate the total distance of points from their cluster mean. Keep the solution with the smallest distance.

(g) Describe two situations (with justification) where you might choose a mixture model over K-means.

[2 marks]

When using a data-type for which it is not obvious how to compute a distance (but for which we can compute a likelihood), or when we have data for which we don't want to be limited to isotropic clusters.