Question 1

(a) You are given a dataset $\{(\boldsymbol{x}_n, t_n)\}_{n=1}^{N}$ where $\boldsymbol{x}_n = (\phi_1(x_n), \phi_2(x_n), \ldots, \phi_p(x_n))$ is a p-dimensional row vector of real-valued features associated with an individual n and t_n is the corresponding real-valued target variable. A **linear regression** model introduces a function $y = f(\boldsymbol{x}; \boldsymbol{w})$ that depends linearly on weight vector $\boldsymbol{w}$. Write out the expression for $f(\boldsymbol{x}; \boldsymbol{w})$. How many real numbers in $\boldsymbol{w}$ does the learning algorithm have to learn.

[4 marks]

(b) Express the learning problem introduced in the previous part in terms of a design matrix A, so that $A\boldsymbol{w} = \boldsymbol{y}$. You must outline what the size of the design matrix is, and what each of its rows and columns represents. Use the column vector of targets $\boldsymbol{t} = (t_1, \ldots, t_N)^T$ to define the mean squared error (MSE) loss function.

[4 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(c) The best fit weights $\widehat{\boldsymbol{w}}$ for the MSE is given by $\widehat{\boldsymbol{w}} = \boldsymbol{A}^+\boldsymbol{t}$, where $\boldsymbol{A}^+$ is the pseudo-inverse

$$\boldsymbol{A}^+ = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T.$$

Show that (i) the sum of the residuals is zero; (ii) the vector of each of the p features $(\phi_j(x_1), \phi_j(x_2), \ldots \phi_j(x_N))$, for $j = 1, \ldots, p$ is orthogonal to the vector of residuals.

[5 marks]

(d) Suppose the design matrix has a singular value decomposition

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \sum_k \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^T$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices of size $(N \times N)$ and $(p \times p)$ respectively and Σ contains non-zero entries σ_i only along the diagonal. Show that $\boldsymbol{A}\boldsymbol{v}_i = \sigma_i\boldsymbol{u}_i$.

[3 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(e) If you rewrite the target vector $\boldsymbol{t}$ and weights $\boldsymbol{w}$ as linear combinations of the columns $\boldsymbol{u}_i$ of $\boldsymbol{U}$ and columns $\boldsymbol{v}_i$ of $\boldsymbol{V}$ as follows:

$$\boldsymbol{w} = \sum_i \alpha_i \boldsymbol{v}_i, \ \boldsymbol{t} = \sum_n \beta_n \boldsymbol{u}_n,$$

show that the MSE loss function reduces to

$$\frac{1}{N}\sum_i (\beta_i - \alpha_i \sigma_i)^2.$$

For what choice of the expansion coefficients α_i of the weight vector $\boldsymbol{w}$ is this loss minimised?

[5 marks]

(f) Suppose the target vector $\boldsymbol{t}$ is corrupted by some $(N \times 1)$ additive noise vector $\boldsymbol{e}$, so that $\boldsymbol{t} + \boldsymbol{e} = \sum_i (\beta_i + \epsilon_i)\boldsymbol{u}_i$ where ϵ_i is small. Show that the existence of small singular values σ_i can lead to large changes in the learnt weights.

[4 marks]

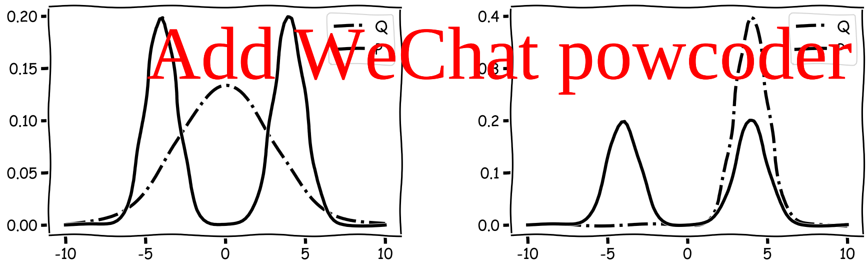
Additional space. Do not use unless necessary. Clearly mark corresponding question.

Question 2

(a) For 2 probability distributions $P_A = \{a_i | a_i \geq 0, \sum_i a_i = 1, i \in \mathcal{X}\}$ and $P_B = \{b_i | b_i \geq 0, \sum_i b_i = 1, i \in \mathcal{X}\}$ over the same event space $\mathcal{X}$ the Kullback-Leibler (KL) divergence $KL(P_A \| P_B)$ are defined as:

$$KL(P_A \| P_B) = \sum_{i \in \mathcal{K}} a_i \log(\frac{a_i}{b_i}).$$

Similar definitions apply in the continuous case. Provide some intuition for when the value of $KL(P_A \| P_B)$ is minimised, motivating how this is exploited in machine learning algorithms. [5 marks]



(b) If the true distribution is given by P, which is bimodal, and you are learning the parameters of a gaussian Q, minimising $KL(P \| Q)$ or $KL(Q \| P)$ can yield either of the two distributions (in dash-dotted lines) shown in the figure. Explain which of the two choices of the KL corresponds to which figure.

[5 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(c) You are given a sequence $\boldsymbol{x} := \{x^{(1)}, \ldots, x^{(N)}\}$ of heads ($x^{(i)} = H$) and tails ($x^{(i)} = T$) which are the outcomes of N tosses of a (potentially biased) coin, with n_H being the number of times heads appears. All possible outcomes of N coin tosses would constitute the event space $\mathcal{X}$. A binomial distribution $B(N, \theta)$ sets the probability of occurrence of n_1 events of type 1, and $n_2 = N - n_1$ of type 2 as

$$P(n_1, n_2 | N, \theta) = \frac{N!}{n_1! n_2!} \theta^{n_1} (1 - \theta)^{n_2}.$$

Describe how you would fit the data $\mathcal{X}$ to a binomial distribution using maximum likelihood estimation (MLE) and find the result $\theta_{MLE} = \frac{n_H}{N}$.

[8 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(d) Use Bayes' theorem to write the posterior probability $P(\theta|\boldsymbol{x})$ of the parameters θ upon observing some data $\boldsymbol{x}$. Discuss how a conjugate Beta prior

$$\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} = \frac{1}{Z}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

introduces “pseudo-counts" to affect the estimation of parameters of the binomial distribution as above, and combats the problem of overfitting, particularly for data sets of small size N. (Z is the normalisation constant.)

[7 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

Question 3

(a) You are given a dataset $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$ where $\boldsymbol{X}$ is a $(N \times p)$ matrix, each row of which $\boldsymbol{x}_n$ is a p-dimensional vector of real-valued features associated with an individual i. Write down what the (i, j)-th components of the matrices $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$ are.

[5 marks]

(b) If the square of the length of $\boldsymbol{x}_n$ defined above is $\|\boldsymbol{x}_n\|^2 = \sum_{i=1}^p x_{n,i}^2$, show that

$$\text{Tr}(\boldsymbol{X}\boldsymbol{X}^T) = \sum_{n=1}^N \|\boldsymbol{x}_n\|^2,$$

where Tr is the matrix trace.

[3 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(c) From the given matrix $\boldsymbol{X}$ as above, write out the (i, j) entry of the covariance matrix of features.

[3 marks]

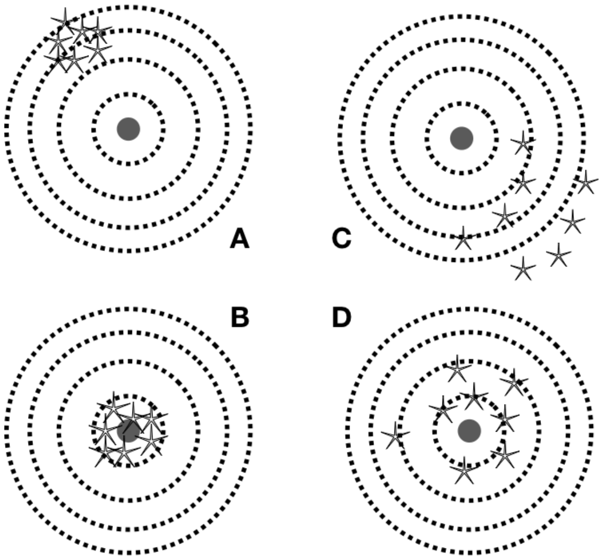
(d) Explain in detail how you would use Principal Components Analysis (PCA) to reduce the dimensionality p of the feature set. You have to introduce the eigenvalues and eigenvectors of the covariance matrix, and explain how they are used to perform this task.

[5 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(e) A uniform random variable $\mathcal{U}(a,b)$ has mean $\frac{a+b}{2}$ and variance $\frac{(b-a)^2}{12}$. For any sample $\boldsymbol{x}$ of N numbers $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$ drawn from a uniform distribution $\mathcal{U}(1,3)$ its average $< \boldsymbol{x} >$ is a random variable. What is the mean and standard deviation of this random variable?

[4 marks]



Additional space. Do not use unless necessary. Clearly mark corresponding question.

(f) The adjacent figure illustrates the performance of a learned machine on a task whose target is at the centre of each set of concentric circles. The stars represents the prediction of the machine on test data. Explain the meaning of the terms bias and variance of a learning method using the figures as a guide.

[5 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

Question 4

(a) Describe the k-means clustering algorithm.

[7 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(b) For a 2-class Gaussian classifier with class label $c = A, B$, in a two dimensional feature space with points $\boldsymbol{X} = (x, y)^T \in \mathbb{R}^2$ the probability distribution functions are

$$p(\boldsymbol{X}|c) = \frac{\sqrt{|\boldsymbol{\Lambda}_c|}}{2\pi} \exp\left(-\frac{1}{2}(\boldsymbol{X} - \boldsymbol{\mu}_c)^T \boldsymbol{\Lambda}_c (\boldsymbol{X} - \boldsymbol{\mu}_c)\right), \; c = A, B.$$

Consider the special case

$$\boldsymbol{\Lambda}_A = \boldsymbol{\Lambda}_B = \boldsymbol{\Lambda} = \begin{pmatrix} a & b \\ b & d \end{pmatrix}, a, b, d > 0, \text{ and means } \boldsymbol{\mu}_A = \boldsymbol{\mu} = -\boldsymbol{\mu}_B.$$

Draw the contours of the two pdfs.

[4 marks]

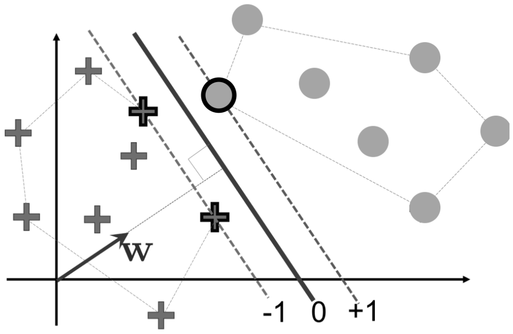
Additional space. Do not use unless necessary. Clearly mark corresponding question.

(c) For the conditional pdfs introduced above, show that the decision boundary $P(A|\boldsymbol{X}) = P(B|\boldsymbol{X})$ for equal priors ($P(A) = P(B)$) is the set of points $\boldsymbol{X} = (x, y)^T$ described by $\boldsymbol{X}^T \boldsymbol{\Lambda} \boldsymbol{\mu} = 0$.

[3 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(d) A linear hyperplane is described by the equation $y = \boldsymbol{w} \cdot \boldsymbol{x} + b$. The decision boundary in the figure is the line (representing a hyperplane) for which $y = 0$ (labelled 0) and is perpendicular to $\boldsymbol{w}$. The two parallel hyperplanes that go through the support vectors (points with thickened edges) are indicated by the values $y = \pm 1$. Explain why a large margin is necessary for robust classification. From the geometry of the figure show that the size of the margin along the direction of $\boldsymbol{w}$ is $\frac{2}{\|\boldsymbol{w}\|}$. You may take $\boldsymbol{x}_+$ and $\boldsymbol{x}_-$ to be support vectors.



[6 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(e) Explain how the max margin classifier for the training set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ is expressed as the solution to the constrained optimisation problem

$$\min_{\boldsymbol{w},b} \max_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_n \alpha_n \left(y_n(\boldsymbol{w}^T\boldsymbol{x}_n + b) - 1\right), \ \alpha_n \geq 0.$$

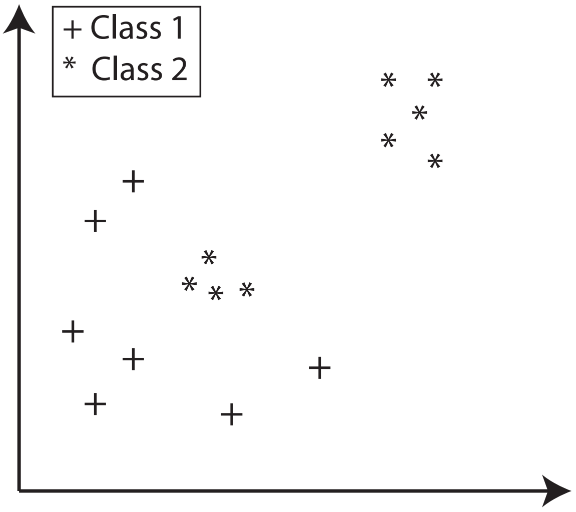
How many constraints are enforced by each Lagrange multiplier α_n and what does each constraint encode?

[5 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

TURN OVER

Question 5 Suppose you are given the 2 class dataset in the figure below.



(a) Would a perceptron be a suitable classifer for this dataset? Briefly explain your answer.

[3 marks]

(b) Would a neural network with 1 hidden layer and no bias be a suitable classifier for this dataset? Briefly explain your answer.

[3 marks]

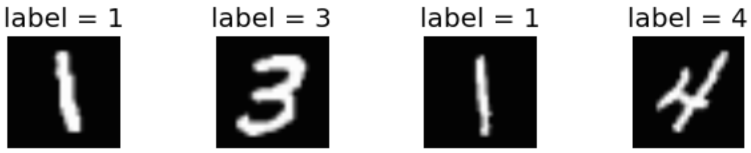
Additional space. Do not use unless necessary. Clearly mark corresponding question.

(c) Assume you now have a fully functional neural network architecture in place and you are training your model. How can you check to see if you are overfitting? Explain exactly what you could plot from your output and what on this plot may indicate overfitting.

[7 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

Now assume you are working with a different dataset of 1000 images. These images are from the MNIST dataset and consist of handwritten digits from 0 to 9. Some samples are given below. Each one of the images consists of 28 x 28 pixels each with a value of 0 or 1 and you want to build a neural network to classify the digits. Note, for any of the answers below, if you do not have a calculator and cannot work out the final number you will not be penalised. Simply write out the simplest form of the equation to the answer of the question.]



(d) How many inputs would you need in your neural network?

[3 marks]

(e) How many neurons would you need in your output layer?

[3 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

(f) For a ReLU function defined as $f^{+}(x, a) := \max(x - a, 0)$, evaluate the output of the network:

$$\begin{aligned} h &= \quad 1 - f^{+}(x_1 + x_2, 0.4) \\ y &= [f^{+}(h, 0.1) - 2f^{+}(h, 0.5) + f^{+}(h, 0.9)] \end{aligned}$$

on the set D of input pairs (x_1, x_2), where

$$D := \{(0.1, 0.2), (0.2, 0.9), (0.8, 0.1), (0.7, 0.9)\}.$$

What pattern has this network captured? You should create a table with a column of inputs on the left and results of the stages of the computation on other columns.

(x_1, x_2)	$\overbrace{x_1 + x_2}^{\text{s}}$	$1 - f^{+}(\text{s}, 0.4)$	$f^{+}(h, 0.1)$	$f^{+}(h, 0.5)$	$f^{+}(h, 0.9)$	y
$(0.1, 0.2)$						
$(0.2, 0.9)$						
$(0.8, 0.1)$						
$(0.7, 0.9)$						

[6 marks]

Additional space. Do not use unless necessary. Clearly mark corresponding question.

END OF PAPER