Guidelines

- You should answer all **TWO** questions.
- Note that not all questions carry equal marks.
- You should submit your final report **as a pdf** via the module's **Moodle** page.
- Within your report you should begin each question on a **new page**.
- You should preface your report with a single page containing, on two lines:
  - The module code and assignment title: **'COMP0036: Assignment 2'**
  - Your candidate number: **'Candidate Number: [YOUR NUMBER]'**
- Your report should be **neat** and **legible**.
- You must use LaTeX to format the report. A template, '`COMP0036_Solution_Template.tex`', is provided on the module's Moodle page for this purpose.
- Please attempt to express your answers as succinctly as possible.
- Please note that if your answer to a question or sub-question is illegible or incomprehensible to the marker then you will receive no marks for that question or sub-question.
- Please remember to detail your working, and state clearly any assumptions which you make.
- Unless a question specifies otherwise, then please make use of the **Notation** section as a guide to the definition of objects.
- Failure to adhere to any of the guidelines may result in question-specific deduction of marks. If warranted these deductions may be punitive, and on occasion may result in no marks being awarded for the assignment.

Notation & Formulae

Inputs:

$\mathbf{x} = [1, x_1, x_2, ..., x_m]^T \in \mathbb{R}^{m+1}$

Outputs:

$y \in \mathbb{R}$ for regression problems

$y \in \{0, 1\}$ for binary classification problems

Training Data:

$\mathcal{S} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{n}$

Input Training Data:

The design matrix, $\mathbf{X}$, is defined as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}^{(1)T} \\ \mathbf{x}^{(2)T} \\ \cdot \\ \cdot \\ \mathbf{x}^{(n)T} \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \cdot & \cdot & x_m^{(1)} \\ 1 & x_1^{(2)} & \cdot & \cdot & x_m^{(2)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_1^{(n)} & \cdot & \cdot & x_m^{(n)} \end{bmatrix}$$

Output Training Data:

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdot \\ \cdot \\ y^{(n)} \end{bmatrix}$$

Data-Generating Distribution:

$\mathcal{S}$ is drawn i.i.d. from a data-generating distribution, $\mathcal{D}$

Marginal & Conditional Distributions of Linear Gaussian Models

Given a Gaussian marginal distribution for $\mathbf{x}$ and a conditional Gaussian distribution for $\mathbf{y}$ given $\mathbf{x}$:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$
$$\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1})$$

Where: $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, $\boldsymbol{\mu} \in \mathbb{R}^n$, $\boldsymbol{\Lambda} \in \mathbb{R}^{n \times n}$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{L} \in \mathbb{R}^{m \times m}$

Then:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T)$$
$$\mathbf{x}|\mathbf{y} \sim \mathcal{N}(\boldsymbol{\Sigma}[\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}], \boldsymbol{\Sigma})$$

Where: $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}$

Properties of Rotation Matrices:

For any n-dimensional rotation matrix, $\mathbf{R} \in \mathbb{R}^{n \times n}$, its properties include:

$\mathbf{R}\mathbf{R}^T = \mathbf{I}$

$\det\mathbf{R} = \pm 1$

1. (a) ***[3 marks]***
State the Representer Theorem, and briefly explain its importance in kernel approaches to machine learning.

(b) ***[4 marks]***
Consider a version of the so-called 'Support Vector Machine' (SVM) classifier optimisation problem:

$$\min_{\mathbf{w}} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{n} \max\left(1 - y^{(i)}\mathbf{w}\cdot\mathbf{x}^{(i)}\right)$$

Here:
$\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^n$ represents a set of training data, where $\mathbf{x} \in \mathbb{R}^m$ are input attributes, while $y \in \{-1, 1\}$ is the output label;
$\mathbf{w} \in \mathbb{R}^m$ is the weight vector of the linear discriminant which we seek, and;
$C > 0$ is some constant.
Is this problem susceptible to the 'Kernel Trick'? Explain.

(c) ***[3 marks]***
State Mercer's Theorem, and briefly explain its importance in kernel approaches to machine learning.

(d) Given: that $\kappa_1 : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$, and $\kappa_2 : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ are valid Mercer kernels; that $\widetilde{\kappa}(\mathbf{u}, \mathbf{v}) = \kappa_1(\mathbf{u}, \mathbf{v})\kappa_2(\mathbf{u}, \mathbf{v})$ is also a valid kernel; that $\alpha > 0$ is some scalar; and that $p(\cdot)$ is a polynomial with non-negative coefficients, then show that the following are valid kernels (here $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$ denote input attribute vectors):

(i) ***[3 marks]***
$\kappa(\mathbf{u}, \mathbf{v}) = \kappa_1(\mathbf{u}, \mathbf{v}) + \kappa_2(\mathbf{u}, \mathbf{v})$

(ii) ***[3 marks]***
$\kappa(\mathbf{u}, \mathbf{v}) = \alpha\kappa_1(\mathbf{u}, \mathbf{v})$

(iii) ***[3 marks]***
$\kappa(\mathbf{u}, \mathbf{v}) = p(\kappa_1(\mathbf{u}, \mathbf{v}))$

(iv) ***[3 marks]***
$\kappa(\mathbf{u}, \mathbf{v}) = \exp(\kappa_1(\mathbf{u}, \mathbf{v}))$

(e) ***[3 marks]***
Hence show that the following 'Radial Basis Function' (RBF) is a valid kernel:

$$\kappa(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{\|\mathbf{u} - \mathbf{v}\|_2^2}{2\sigma^2}\right)$$

Where $\sigma > 0$

[Total for Question 1: 25 marks]

2. We assume an unlabelled dataset, $\left\{\mathbf{x}^{(i)} \in \mathbb{R}^m\right\}_{i=1}^n$, with sample mean, $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}^{(i)}$, and we define $\mathbf{X} = [(\mathbf{x}^{(1)} - \overline{\mathbf{x}}), (\mathbf{x}^{(2)} - \overline{\mathbf{x}}), \ldots, (\mathbf{x}^{(n)} - \overline{\mathbf{x}})]^T$.

   We seek some d-dimensional subspace, where $d < m$, characterised by the following latent variable model:

   Each data point, $\mathbf{x}$, in the dataset has an unknown latent variable, $\mathbf{z} \in \mathbb{R}^d$, associated with it, corresponding to its position in the latent subspace. This variable is the outcome of a Gaussian random variable, $\boldsymbol{\mathcal{Z}}$, such that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

   Conditional on the subspace variable, each $\mathbf{x}$ is the outcome of a Gaussian random variable, $\boldsymbol{\mathcal{X}}$, such that $\mathbf{x}|\mathbf{z} \sim \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \boldsymbol{\Phi})$.

   Here the columns of $\mathbf{W} \in \mathbb{R}^{m\times d}$ define the directions of the subspace which we seek, $\boldsymbol{\mu} \in \mathbb{R}^m$, and $\boldsymbol{\Phi} \in \mathbb{R}^{m\times m}$ is a positive definite covariance matrix.

   We can characterise the model more succinctly as follows:

$$\begin{aligned} \mathbf{x} &= \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \\ \mathbf{z} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}) \end{aligned} \tag{1}$$

   Here $\boldsymbol{\varepsilon}$ is the outcome of a random variable $\boldsymbol{\epsilon}$.

   Furthermore, $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\epsilon}$ are uncorrelated.

   (a) ***[7 marks]***
   Demonstrate that the following model is entirely equivalent to model (1).
   Here $\widetilde{\mathbf{W}} \in \mathbb{R}^{m\times d}$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is a positive definite covariance matrix, with eigenvectors given by the columns of some matrix $\mathbf{Q} \in \mathbb{R}^{d\times d}$ and with associated eigenvalues, $\{\lambda_i\}_{i=1}^d$, which are stored in some matrix $\boldsymbol{\Lambda} = \text{diag}[\lambda_1, \ldots, \lambda_d]$.
   Furthermore, $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\epsilon}$ are uncorrelated:

$$\begin{aligned} \mathbf{x} &= \widetilde{\mathbf{W}}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \\ \mathbf{z} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \\ \boldsymbol{\varepsilon} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}) \end{aligned}$$

   In so doing express $\widetilde{\mathbf{W}}$ in terms of $\mathbf{W}$, $\boldsymbol{\Lambda}$ and $\mathbf{Q}$.

   (b) ***[7 marks]***
   Returning to our original model, (1), demonstrate that the log likelihood of the data, $\ln \mathbb{P}\left(\left\{\mathbf{x}^{(i)} \in \mathbb{R}^m\right\}_{i=1}^n\right)$, under the maximum likelihood assumption, can be expressed as:

$$-\frac{n}{2}\ln|2\pi\mathbf{C}| - \frac{n}{2}\text{tr}\left(\mathbf{C}^{-1}\mathbf{S}\right)$$

   In so doing derive expressions for $\mathbf{C}$ and $\mathbf{S}$ in terms of $\mathbf{X}$, $\mathbf{W}$, $\boldsymbol{\Phi}$.

   (c) ***[5 marks]***
   Let us assume that we have found a maximum likelihood estimator of the matrix $\mathbf{W}$, which we term $\mathbf{W}_{\mathsf{MLE}}$. Demonstrate that the log likelihood derived in part (b) is invariant to rotations of the directions of the latent subspace defined by the columns of

$\mathbf{W}_{\mathsf{MLE}}$, for rotations which take place in the latent space itself. In other words, show that the log likelihood is invariant to the following transformation:

$$\mathbf{W}_{\mathsf{MLE}} \longleftarrow \mathbf{W}_{\mathsf{MLE}}\mathbf{R}$$

Where $\mathbf{R} \in \mathbb{R}^{d\times d}$ is a rotation matrix.

(d) ***[6 marks]***

Now, if we set $\mathbf{\Phi} = \sigma^2\mathbf{I}_m$ for some scalar σ, then it is possible to demonstrate that the log likelihood solution is invariant if we transform the data and the directions of the subspace as follows:

$$\mathbf{x}^{(i)} \longleftarrow \widehat{\mathbf{x}}^{(i)} = \mathbf{U}\mathbf{x}^{(i)} \quad \forall i$$
$$\mathbf{W}_{\mathsf{MLE}} \longleftarrow \widehat{\mathbf{W}}_{\mathsf{MLE}} = \mathbf{U}\mathbf{W}_{\mathsf{MLE}}$$

Where $\mathbf{U} \in \mathbb{R}^{m\times m}$ is some orthogonal matrix.

Let us call this specialisation of model (1), 'model (A)'.

Furthermore, if we set $\mathbf{\Phi} = \mathrm{diag}[\sigma_1^2, \sigma_2^2, ..., \sigma_m^2]$ for scalars $\{\sigma_i\}_{i=1}^m$, then it is possible to demonstrate that the log likelihood solution is invariant if we transform the data, the directions of the subspace, and the conditional noise as follows:

$$\mathbf{x}^{(i)} \longleftarrow \widehat{\mathbf{x}}^{(i)} = \mathbf{D}\mathbf{x}^{(i)} \quad \forall i$$
$$\mathbf{W}_{\mathsf{MLE}} \longleftarrow \widehat{\mathbf{W}}_{\mathsf{MLE}} = \mathbf{D}\mathbf{W}_{\mathsf{MLE}}$$
$$\mathbf{\Phi} \longleftarrow \widehat{\mathbf{\Phi}} = \mathbf{D}\mathbf{D}\mathbf{\Phi}$$

Where $\mathbf{D} \in \mathbb{R}^{m\times m}$ is some diagonal matrix.

Let us call this specialisation of model (1), 'model (B)'.

If we are concerned about issues in the scaling of the dataset then explain whether we should use model (A) or model (B) in order to effect dimensionality reduction.

[Total for Question 2: 25 marks]