Assignment 1: Foundations, Linear Methods, Kernel Techniques

Submission Instructions: Please submit your answers as a single PDF file. Name the file **firstname.lastname.pdf**, then upload it to the iCorsi website before the deadline. Handwriting and scanned documents are not allowed. Late submissions will not be accepted.

The deadline is on October 22 2025.

For questions, please send an email to guillaume.toussaint@usi.ch.

1 Ridge Regression (15 points)

a) Provide the formulation of the ridge regression model.

b) Assuming a set of training data $\mathcal{O} = \{\mathbf{x}_n, t_n\}_{n=1,\dots,N}$, write down the Mean Squared Error loss. Then, based on a), show how to derive the optimal regression parameters.

c) Explain the effect of the regularization parameter λ in ridge regression. What happens when λ is very small? What happens when λ is very large?

d) Given the following dataset with $\lambda = 0.8$:

Observation	X_1	X_2	Y
1	1	2	10
2	2	1	12
3	3	3	18
4	4	2	21

A student claims that the optimal regression estimator is defined by $\hat{y} = \theta_0 + X_1\theta_1 + X_2\theta_2$, with $\theta_0 = 1$, $\theta_1 = 2$, $\theta_2 = 1$.

i. Based on b), how can you demonstrate that these parameters are not optimal?

ii. What are the actual optimal parameters? Compute them. (Show each step)

iii. Compute the MSE for both the claimed parameters and the actual optimal parameters to quantitatively show the difference in fit quality.

2 Kernel Methods (5 points)

a) Explain how kernel methods extend linear algorithms to solve non-linear problems.

b) Given the function $k(x, y) = (x^T y)^d$ with $d \geq 1$, is this a valid kernel? Justify your answer.

c) Compare the memory requirements of kernel ridge regression and non-kernelized linear ridge regression. Which one scales worse with the number of data points, and why?

3 Support Vector Machines (10 points)

Consider the following 2D dataset for binary classification:

Class	x_1	x_2
+	2	3
+	3	1
+	1	3
-	-1	0
-	-2	1
-	1	-1

a) Plot these points. Are the classes linearly separable?

b) Determine the maximum margin hyperplane by inspection. Provide the weight vector $\mathbf{w}$ and bias term b. Identify which points are the support vectors and explain why they qualify as such.

c) Calculate the width of the margin for your proposed solution.

d) Suppose that we apply a polynomial kernel of degree 2 to the dataset. Would the data be linearly separable? Do you expect the margin width to change? Justify your answer.

e) If we added a new point at (2, 0.5) with class +, is a SVM computed on the original dataset a valid solution? Why or why not? How would adding slack variables (*soft-margin SVM*) handle this situation?