1 "Question de cours" (15 points)

1.1 Regression

We want to predict $Y_i \in \mathbb{R}$ as a function of $X_i \in \mathbb{R}$. We consider the following models:

(a) Linear regression
(b) 2-nd order polynomial regression
(c) 10-th order polynomial regression
(d) Kernel ridge regression with a Gaussian kernel
(e) $k$-nearest neighbor regression

We consider the following regression problems.

Answer each of the following questions with no justification.

1. (1 point) If $Y \in \mathbb{R}^n$ is the output vector and $X \in \mathbb{R}^n$ is the input vector. Write the expression of the estimator for linear regression.

2. (3 points) What are the time and space complexities
   - in $n$ and $d$ of $d$-th order polynomial regression,
   - in $n$ of kernel ridge regression,
   - in $n$ and $k$ of $k$-nearest neighbor regression?
3. (1 point) What are the hyper-parameters of kernel ridge regression and \( k \)-nearest neighbors?

4. (2.5 points) For each problem, what would be the good model(s) to choose? (no justification)

5. (1 point) What models would lead to over-fitting in Problem 1.

6. (1 point) Provide one solution to deal with over-fitting.

1.2 Classification

We aim at predicting \( Y_i \in \{0, 1\} \) as a function of \( X_i \in \mathbb{R}^2 \) (with the notation \( \circ = 0 \) and \( \times = 1 \)). We consider the following models:

(a) Logistic regression
(b) Linear discriminant analysis
(c) Logistic regression with 2-nd order polynomials
(d) Logistic regression with 10-th order polynomials
(e) \( k \)-nearest neighbor classification

We consider the following classification problems.

![Classification 1](image1.png) ![Classification 2](image2.png) ![Classification 3](image3.png) ![Classification 4](image4.png) ![Classification 5](image5.png)

Answer each of the following questions with no justification.

7. (2 points) Write the optimization problem that logistic regression is solving. How is it solved?

8. (1 point) What is the main assumption on the data distribution made by linear discriminant analysis?

9. (2.5 points) For each problem, what would be the good model(s) to choose? (no justification)

2 Gradient descent (GD) (7 points)

The goal of this exercise is to study GD with a constant step-size in the simplest setting. We consider a strictly convex quadratic function \( f : \mathbb{R}^d \to \mathbb{R} \) of the form

\[
f(\theta) = \frac{1}{2} \theta^\top H \theta - g^\top \theta.
\]

10. (1 point) What conditions on \( H \) lead to a strictly convex function? Compute a minimizer \( \theta_* \) of \( f \). Is it unique?

11. (1 point) We consider the gradient descent recursion:

\[
\theta_t = \theta_{t-1} - \gamma f'(\theta_{t-1}).
\]

What is the expression of \( \theta_t - \theta_* \) as a function of \( \theta_{t-1} - \theta_* \), and then as a function of \( \theta_0 - \theta_* \)?
12. (1 point) Compute \( f(\theta) - f(\theta^*) \) as a function of \( H \) and \( \theta - \theta^* \).

13. (2 points) Assuming a lower-bound \( \mu > 0 \) and upper-bound \( L \) on the eigenvalues of \( H \), and a step-size \( \gamma \leq 1/L \), show that for all \( t > 0 \),
\[
f(\theta_t) - f(\theta^*) \leq (1 - \gamma \mu)^{2t} \left[ f(\theta_0) - f(\theta^*) \right].
\]

What step-size would be optimal from the result above?

14. (2 points) Only assuming an upper-bound \( L \) on the eigenvalues of \( H \), and a step-size \( \gamma \leq 1/L \), show that for all \( t > 0 \),
\[
f(\theta_t) - f(\theta^*) \leq \frac{\|\theta_0 - \theta^*\|^2}{8\gamma t}.
\]

What step-size would be optimal from the result above?

15. (Removed)

3 Between K-means and PCA (7 points)

We consider the data \( x_1, \ldots, x_n \in \mathbb{R}^2 \), as follows:

![Data Plot](image)

16. (1 point) Why are K-means and principal component analysis not suited to represent the data?

17. (1 point) We want (1) to cluster the data into \( K \) disjoint groups represented by a partition \( A = (A_1, \ldots, A_K) \) of \( \{1, \ldots, n\} \), and (2) to find one-dimensional affine subspaces represented by vectors \( \mu_k \in \mathbb{R}^d \) and \( \Delta_k \in \mathbb{R}^d \), \( k \in \{1, \ldots, K\} \) such that
\[
J(A, \mu, \Delta) = \sum_{k=1}^K \sum_{i \in A_k} d(x_i, \mu_k + \mathbb{R}\Delta_k)^2 = \sum_{k=1}^K \sum_{i \in A_k} \inf_{t_i \in \mathbb{R}} \|x_i - \mu_k - t_i \Delta_k\|_2^2
\]
is minimal. What does the situation where all \( \Delta_k \) are equal to zero correspond to? What does the situation where \( K = 1 \) correspond to (no justification required)?

18. (2 points) We consider an alternating minimization algorithm. How can we minimize with respect to the partition \( A \) when \( \mu, \Delta \) are fixed?

19. (2 points) How to minimize with respect to \((\mu, \Delta)\) when \( A \) is fixed (one can start for each of the \( K \) clusters to jointly minimize in \( \mu_k \) and \( t_i \), for \( i \in A_k \)).

3
20. (1 point) What are the convergence properties of this algorithm?

4 SVM – Support vector machine classifier (15 points)

Let $D_n = \{(x_i, y_i)\}_{1 \leq i \leq n} \in (\mathbb{R}^d \times \{-1, 1\})^n$ be a data set of $n$ observations. The goal of this exercise is to study the SVM classifier $\hat{w}$ defined for $\lambda > 0$

\[
\hat{w} \in \arg\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( 1 - y_i w^\top x_i \right)_+ + \lambda \|w\|^2 \right\},
\]

where $(\cdot)_+ = \max\{\cdot, 0\}$.

21. (1 point) Show that the minimum is attained at a unique point.

22. (2 points) Prove that $\hat{w}$ is a linear combination of $x_1, \ldots, x_n$.

23. (1 point) Prove that $\hat{w} = \sum_{j=1}^{n} \hat{\beta}_j x_j$ where $\hat{\beta} \in \mathbb{R}^n$ is the solution to

\[
\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( 1 - y_i (K\beta)_i \right)_+ + \lambda \beta^\top K\beta \right\},
\]

where $K = [x_i^\top x_j]_{1 \leq i, j \leq n} \in \mathbb{R}^{n \times n}$ is the Gram matrix.

24. (1 point) Show that this minimization problem is equivalent to

\[
\hat{\beta} \in \arg\min_{\beta, \xi \in \mathbb{R}^n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \beta^\top K\beta \right\}.
\]

25. (2 points) From the KKT conditions, check that $\hat{\beta}_i = \frac{y_i \hat{\alpha}_i}{2n}$ with $\hat{\beta}$ and $\hat{\alpha}$ satisfying

\[
\hat{\alpha}_i (y_i (K\hat{\beta})_i - (1 - \hat{\xi}_i)) = 0 \quad \text{and} \quad \frac{1}{n} - \hat{\alpha}_i \hat{\xi}_i = 0.
\]

26. (3 points) Prove the following properties

- if $y_i \hat{w}^\top x_i > 1$ then $\hat{\beta}_i = 0$
- if $y_i \hat{w}^\top x_i < 1$ then $\hat{\beta}_i = \frac{y_i}{2n}$
- if $y_i \hat{w}^\top x_i = 1$ then $0 \leq \hat{\beta}_i y_i \leq \frac{1}{2n}$.

27. (1 point) Give a geometric interpretation of this result

28. (1 point) Does strong duality hold?

29. (2 points) Prove that $\hat{\alpha}_i$ is the solution to the dual problem

\[
\hat{\alpha} \in \arg\max_{0 \leq \alpha_i \leq \frac{1}{2}} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{4\lambda} \sum_{i=1}^{n} \sum_{j=1}^{n} K_{i,j} y_i y_j \alpha_i \alpha_j \right\}.
\]

30. (1 point) How would you adapt this estimator to non-linear classification rules?