Mosig: "From Basic Machine Learning models to Advanced Kernel Learning"
Final exam

Pierre Gaillard, Michael Arbel and Julien Mairal

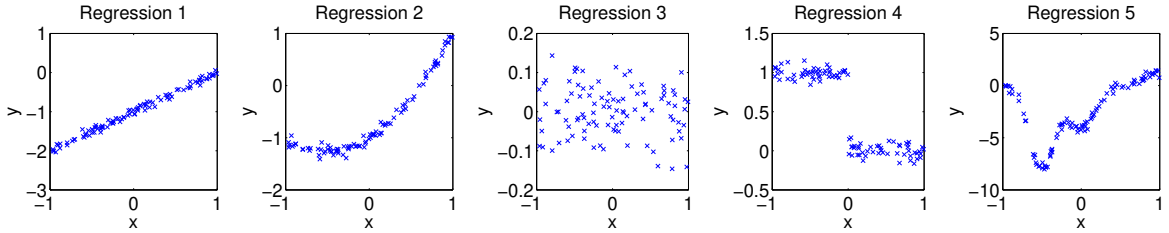
General instructions. No material (slides, book, laptop, cell phone, ...) is allowed for the final written exam. The exercises are tentatively ordered from easiest to most difficult ones.

Exercice 1. Basic machine learning models

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 (with linear features)
(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. 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 (with linear feature map).

   Solution In one dimension, the estimator of linear regression solves the following optimization problem:

   $$\hat{\beta}_n \in \operatorname*{argmin}_{\beta \in \mathbb{R}^2} \left\| Y - \beta_0 + X\beta_1 \right\|^2.$$

   Solving the gradients yields: $\hat{\beta}_0 = \bar{Y}$ where $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\hat{\beta}_1 = (X^\top X)^{-1} X^\top (Y - \bar{Y})$. Another solution is to add the intercept in the input matrix, writing $\tilde{X} := \big[1, X\big]$ the $(n \times 2)$ matrix where the first column is $(1, \dots, 1)^\top \in \mathbb{R}^n$, we have $\hat{\beta}_n = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top Y$. ∎

2. 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?

   Solution Polynomial regression for one-dimensional inputs needs to compute $(Z^\top Z)^{-1} Z^\top X$ where $Z = [1, X, X^2, \dots, X^d]$ is an $(n \times (d+1))$ matrix. The matrix multiplication $Z^\top Z$ costs $O(nd^2)$ and the matrix inversion of the $(d+1) \times (d+1)$ matrix $(Z^\top Z)^{-1}$ costs $O(d^3)$ time.

Kernel regression needs to compute $\alpha = (K_{nn} + n\lambda I_n)^{-1} Y \in \mathbb{R}^n$, where $K_{nn} = \left[k(x_i, x_j)\right]_{1 \leq i,j \leq n}$. For a new input $x \in \mathbb{R}$, it then predicts $\hat{f}_\lambda(x) = \sum_{i=1}^n k(x_i, x)\alpha_i$. The algorithm thus needs to inverse the $n \times n$ matrix $K_{nn} + n\lambda I_n$ which requires $O(n^3)$ time and $O(n^2)$ space.

The k-NN regression does not need any training. The time complexity of the training part is thus $O(1)$, while for space it only needs to store all points which requires $O(n)$. However, a naive implementation of k-NN (there are optimized versions using trees) requires $O(nk)$ runtime to make a prediction.

Regression model	Time complexity	Space complexity
Polynomial regression	$O(d^3 + d^2 n)$	$O(d^2 + dn)$
Kernel ridge regression	$O(n^3)$	$O(n^2)$
k-nearest neighbor	$O(nk)$ (for prediction)	$O(n)$

■

3. What are the hyper-parameters of kernel ridge regression and k-nearest neighbors?

   Solution Kernel ridge regression with a Gaussian kernel requires two hyper-parameters: the regularization parameter $\lambda > 0$ and the bandwidth of the Gaussian kernel: $\sigma > 0$.

   k-nearest neighbor only needs the number of neighbors $k \geq 1$. ■

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

   Solution Several solutions are possible for each problem. We only choose here the ones that seem to be the most appropriate (i.e., the simplest one). Some methods such as kernel ridge regression would need however to be regularized enough.

Problem	1	2	3	4	5
Best models among (a)-(e)	a	b	No method will perform well. The best would be (a) to avoid over-fitting.	e	c, d

■

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

   Solution In problem 1, the relation between X and Y seem to be linear. Models c, d, and e might lead to over-fitting (though there seems to be sufficiently many points in the dataset) if they are not regularized enough. ■

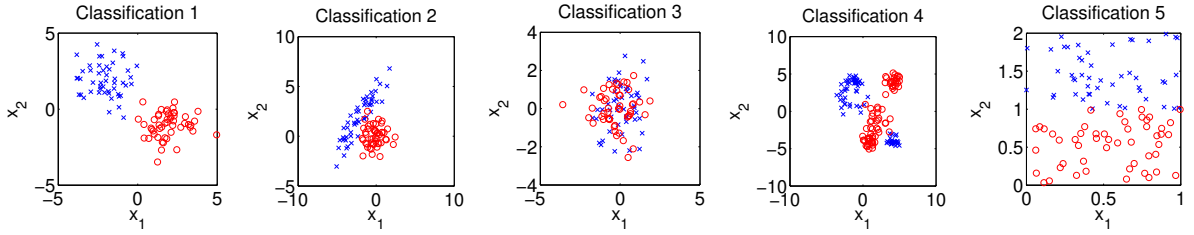
6. Provide one solution to deal with over-fitting.

   Solution A solution is to use cross-validation to calibrate the hyper-parameters to regularize enough the methods (such as the regularization parameter λ in KRR or the bandwidth σ). Cross-validation can also be used to select the best model among (a-e). ■

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) Linear Logistic regression
(b) Multi-Layer Perceptron
(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.



Answer each of the following questions with no justification.

7. Write the optimization problem that logistic regression is solving. How is it solved?

   Solution Logistic regression solves the following optimization problem:

$$\min_{\beta_0,\beta\in\mathbb{R}^2} \sum_{i=1}^{n} \ell(\beta_0 + \beta^\top X_i, Y_i)$$

   where $\beta_0 \in \mathbb{R}$ is the intercept (which may be included into the inputs) and $\ell(\hat{y}, y) = y \log\left(1 + e^{-\hat{y}}\right) + (1-y)\log\left(1 + e^{\hat{y}}\right)$ is the logistic loss. Contrary to least square regression, there is no closed form solution. One needs thus to use iterative convex optimization algorithms such as Newton's method or gradient descent. ■

8. Write the update rule of a single unit of a Multilayer Perceptron. What activation function would you choose/not choose for these problems?

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

   Solution Again several solutions are possible for each problem. We only choose here the ones that seem to be the most appropriate (i.e., the simplest one).

Problem	1	2	3	4	5
Best models among (a)-(e)	a,b,e	c	No model will be good. Choose the simplest to avoid over-fitting: a,b,e	d, e	a

■

Exercice 2. Kernel k-means

In order to cluster a set of vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ into K groups, we consider the minimization of:

$$C(z, \mu) = \sum_{i=1}^{n} \| x_i - \mu_{z_i} \|^2$$

over the cluster assignment variable z_i (taking values in $1, \ldots, K$ for all $i = 1, \ldots, n$) and over the cluster means $\mu_i \in \mathbb{R}^p, i = 1, \ldots, K$.

1. Starting from an initial assignment z^0, we can try to minimize $C(z, \mu)$ by iterating:

$$\mu^i = \operatorname*{argmin}_{\mu} C(z^i, \mu)\,, \qquad z^{i+1} = \operatorname*{argmin}_{z} C(z, \mu^i)\,.$$

   Explicit how both minimization can be carried out (note: this method is called k-means).

2. Propose a similar iterative algorithm to perform k-means in the RKHS $\mathcal{H}$ of a p.d. kernel K over $\mathbb{R}^p$, i.e., to minimize:

$$C_K(z, \mu) = \sum_{i=1}^{n} \| \Phi(x_i) - \mu_{z_i} \|^2 ,$$

where $\Phi : \mathbb{R}^p \to \mathcal{H}$ satisfies $\Phi(x)^\top \Phi(x') = K(x, x')$.

3. Let Z be the $n \times K$ assignment matrix with values $Z_{ij} = 1$ if x_i is assigned to cluster j, 0 otherwise. Let $N_j = \sum_{i=1}^{n} Z_{ij}$ be the number of points assigned to cluster j, and L be the $K \times K$ diagonal matrix with entries $L_{ii} = 1/N_i$. Show that minimizing $C_K(z, \mu)$ is equivalent to maximizing over the assignment matrix Z the trace of $L^{1/2} Z^\top K Z L^{1/2}$.

4. Let $H = ZL^{1/2}$. What can we say about $H^\top H$? Do you see a connection between kernel k-means and kernel PCA? Propose an algorithm to estimate Z from the solution of kernel PCA.

Exercice 3. RKHS of a inner product of features

Let $\mathcal{X}$ be a set and $\mathcal{F}$ be a Hilbert space. Let $\Psi : \mathcal{X} \to \mathcal{F}$, and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be:

$$\forall x, x' \in \mathcal{X}, \quad K(x, x') = \langle \Psi(x), \Psi(x') \rangle_{\mathcal{F}} .$$

1. Show that K is a positive definite kernel on $\mathcal{X}$.

2. Consider the set $\mathcal{H}$ defined as follow:

$$\mathcal{H} = \{ f : \mathcal{X} \to \mathbb{R} \quad | \quad (\exists w \in \mathcal{F})(\forall x \in \mathcal{X}) \qquad f(x) = \langle w, \Psi(x) \rangle_{\mathcal{F}} \} \tag{1}$$

and define the map T as follows:

$$T : \mathcal{F} \to \mathcal{H} \tag{2}$$

$$w \mapsto f_w := \langle w, \Psi(.) \rangle_{\mathcal{F}}. \tag{3}$$

Show that T has a closed null-space (i.e. if a sequence w_n in $\mathcal{F}$ satisfying $T(w_n) = 0$ converges to an element $w \in \mathcal{F}$ and then $T(w) = 0$).

3. Show that there exists a closed sub-space V of $\mathcal{F}$ on which the restriction of T (denoted T_V) is an isomorphism between V and $\mathcal{H}$. (Hint: recall that any closed sub-space of a Hilbert space admits a unique orthogonal supplement.)

4. Consider the following bilinear form defined on $\mathcal{H}$:

$$\langle f, g \rangle_{\mathcal{H}} = \langle T_V^{-1}(f), T_V^{-1}(g) \rangle_{\mathcal{F}} \tag{4}$$

Show that $\langle f, g \rangle_{\mathcal{H}}$ is an inner-product on $\mathcal{H}$.

5. Deduce that $\mathcal{H}$ is Hilbert space endowed with inner-product $\langle ., . \rangle_{\mathcal{H}}$.

6. Show that $\mathcal{H}$ is an RKHS with kernel K.

7. Show that for any $f \in \mathcal{H}$, the following equality holds:

$$\|f\|_{\mathcal{H}} = \inf_{\substack{w \in \mathcal{F} \\ f = T(w)}} \|w\|_{\mathcal{F}}. \tag{5}$$

8. Consider the space:

$$\mathcal{H}' := \{f : \mathcal{X} \to \mathbb{R} \quad | \quad f(x) = \langle w, \Psi(x) \rangle_{\mathcal{H}}, \quad s.t. \quad w \in \overline{Span(Im(\Psi))}\}. \tag{6}$$

where $Im(\Psi)$ stands for the image set of $\mathcal{X}$ by Ψ and for any subset A of $\mathcal{F}$, $\overline{Span(A)}$ denote the closure in $\mathcal{F}$ of the vector space obtained by finite linear combination of elements in A. Show that $\mathcal{H}'$ is equal to $\mathcal{H}$.