Final Exam
Introduction to Statistical Learning
ENS 2018-2019

January 25th 2019

The duration of the exam is 3 hours. You may use any printed references including books. **The use of any electronic device** (computer, tablet, calculator, smartphone) is **forbidden**.

All questions require a proper mathematical justification or derivation (unless otherwise stated), but most questions can be answered concisely in just a few lines. No question should require lengthy or tedious derivations or calculations.

Answers can be written in French or English.

1 "Question de cours" (16 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

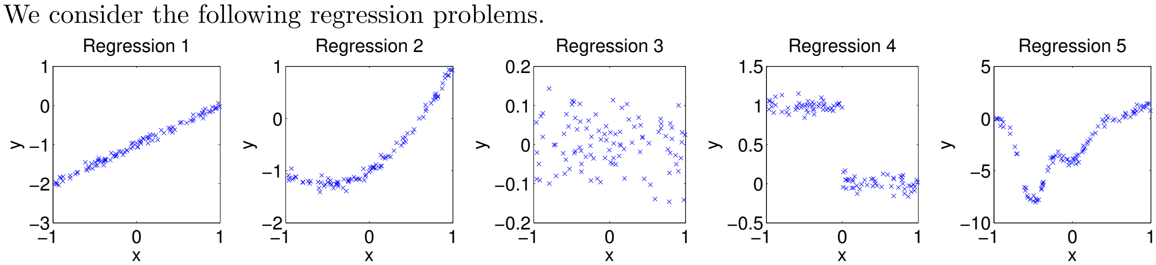
(b) 2-nd order polynomial regression

(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. (2 points) 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.

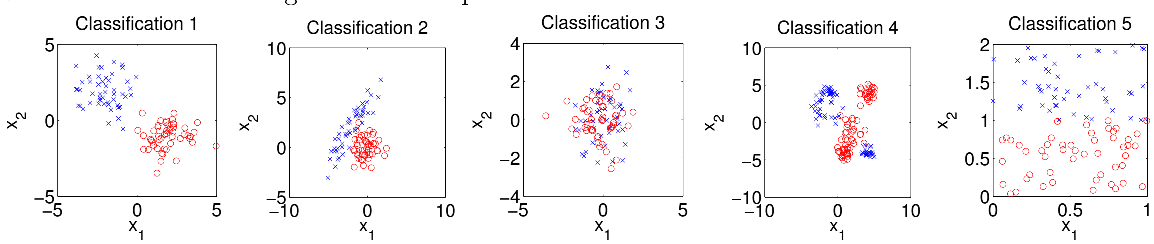
6. (1 point) Provide one solution to deal with over-fitting.

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.



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 Projection onto the ℓ_1-ball (13 points)

Let $z \in \mathbb{R}^n$ and $\mu \in \mathbb{R}_+^*$. We consider the following optimization problem:

$$\text{minimize } \frac{1}{2}\|x - z\|_2^2 \text{ with respect to } x \in \mathbb{R}^n \text{ such that } \|x\|_1 \leqslant \mu.$$

10. (1 point) Show that the minimum is attained at a unique point.

11. (1 point) Show that if $\|z\|_1 \leqslant \mu$, the solution is trivial.

12. (2 points) We now assume $\|z\|_1 > \mu$. Show that the minimizer x is such that $\|x\|_1 = \mu$.

13. (2 points) Show that the components of the solution x have the same signs as the ones of z. Show then that the problem of orthogonal projection onto the ℓ_1-ball can be solved from an orthogonal projection onto the simplex, for some well-chosen u, that is:

$$\text{minimize } \frac{1}{2}\|y-u\|_2^2 \text{ with respect to } y \in \mathbb{R}_+^n \text{ such that } \sum_{i=1}^n y_i = 1.$$

14. (3 points) Using a Lagrange multiplier β for the constraint $\sum_{i=1}^n y_i = 1$, show that a dual problem may be written as follows:

$$\text{maximize } -\frac{1}{2}\sum_{i=1}^n \max\{0, u_i - \beta\}^2 + \frac{1}{2}\|u\|_2^2 - \beta \text{ with respect to } \beta \in \mathbb{R}.$$

Does strong duality hold?

15. (4 points) Show that the dual function is continuously differentiable and piecewise quadratic with potential break points at each u_i, and compute its derivative at each break point. Describe an algorithm for computing β and y with complexity $O(n \log n)$.

3 Stochastic gradient descent (SGD) (23 points)

The goal of this exercise is to study SGD 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

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

16. (1 point) What conditions on H lead to a strictly convex function? Compute a minimizer θ_* of f. Is it unique?

17. (2 points) 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 $\theta_0 - \theta_*$?

18. (1 point) Compute $f(\theta) - f(\theta_*)$ as a function of H and $\theta - \theta_*$.

19. (2 points) Assuming a lower-bound $\mu > 0$ and upper-bound L on the eigenvalues of H, and a step-size $\gamma \leqslant 1/L$, show that for all $t > 0$,

$$f(\theta_t) - f(\theta_*) \leqslant (1-\gamma\mu)^{2t}\big[f(\theta_0) - f(\theta_*)\big].$$

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

20. (2 points) Only assuming an upper-bound L on the eigenvalues of H, and a step-size $\gamma \leqslant 1/L$, show that for all $t > 0$,

$$f(\theta_t) - f(\theta_*) \leqslant \frac{\|\theta_0 - \theta_*\|^2}{8\gamma t}.$$

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

21. (2 points) We consider the stochastic gradient descent recursion:

$$\theta_t = \theta_{t-1} - \gamma\big[f'(\theta_{t-1}) + \varepsilon_t\big],$$

where ε_t is a sequence of independent and identically distributed random vectors, with zero mean $\mathbb{E}(\varepsilon_t) = 0$ and covariance matrix $C = \mathbb{E}(\varepsilon_t \varepsilon_t^\top)$.

What is the expression of $\theta_t - \theta_*$ as a function of $\theta_{t-1} - \theta_*$ and ε_t, and then as a function of $\theta_0 - \theta_*$ and all $(\varepsilon_k)_{k \leqslant t}$?

22. (2 points) Compute the expectation of θ_t and relate it to the (non stochastic) gradient descent recursion.

23. (3 points) Show that

$$\mathbb{E} f(\theta_t) - f(\theta_*) = \frac{1}{2}(\theta_0 - \theta_*)^\top H (I - \gamma H)^{2t}(\theta_0 - \theta_*) + \frac{\gamma^2}{2} \operatorname{tr} CH \sum_{k=0}^{t-1} (I - \gamma H)^{2k}.$$

24. (2 points) Assuming that $\gamma \leqslant 1/L$ (where L is an upper-bound on the eigenvalues of H), show that $H \sum_{k=0}^{t-1} (I - \gamma H)^{2k} = \frac{1}{\gamma}(2 - \gamma H)^{-1}(I - (I - \gamma H)^{2t})$, and that its eigenvalues are all between 0 and $1/\gamma$.

25. (2 points) Assuming a lower-bound $\mu > 0$ and upper-bound L on the eigenvalues of H, and a step-size $\gamma \leqslant 1/L$, show that for all $t > 0$,

$$\mathbb{E} f(\theta_t) - f(\theta_*) \leqslant (1 - \gamma\mu)^{2t}\big[f(\theta_0) - f(\theta_*)\big] + \frac{\gamma}{2} \operatorname{tr} C.$$

26. (4 points) Only assuming an upper-bound L on the eigenvalues of H, and a step-size $\gamma \leqslant 1/L$, show that for all $t > 0$,

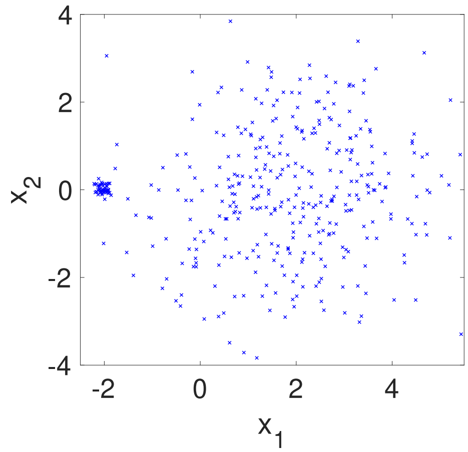
$$\mathbb{E} f(\theta_t) - f(\theta_*) \leqslant \frac{\|\theta_0 - \theta_*\|^2}{8\gamma t} + \frac{\gamma}{2} \operatorname{tr} C.$$

Considering that t is known in advance, what would be the optimal step-size from the bound above? Comment on the obtained bound with this optimal step-size.

4 Mixture of Gaussians (24 points)

In this exercise, we consider an unsupervised method that improves on some shortcomings of the K-means clustering algorithm.

27. (1 point) Given the data below, plot (roughly) the clustering that K-means with $K = 2$ would lead to.



We consider a probabilistic model on two variables X and Z, where $X \in \mathbb{R}^d$ and $Z \in \{1, \ldots, K\}$. We assume that

(a) the marginal distribution of Z is defined by the vector in the simplex $\pi \in \mathbb{R}^K$ (that is with non-negative components which sum to one) so that $\mathbb{P}(Z = k) = \pi_k$,

(b) the conditional distribution of X given $Z = k$ is a Gaussian distribution with mean μ_k and covariance matrix $\sigma_k^2 I$.

28. (1 point) Write down the log-likelihood $\log p(x, z)$ of a single observation $(x, z) \in \mathbb{R}^d \times \{1, \ldots, K\}$.

29. (3 points) We assume that we have n independent and identically distributed observations (x_i, z_i) of (X, Z) for $i = 1, \ldots, n$. Write down the log likelihood of these observations, and show that it is a sum of a function of π and a function of $(\mu_k, \sigma_k)_{k \in \{1,\ldots,K\}}$.

    It will be useful to introduce the notation $\delta(z_i = k)$, which is equal to one if $z_i = k$ and 0 otherwise, and double summations of the form $\sum_{k=1}^K \sum_{i=1}^n \delta(z_i = k) J_{ik}$ for a certain J.

30. (4 points) In the setting of the question above, what are the maximum likelihood estimators of all parameters?

31. (2 points) Show that the marginal distribution on X has density

$$p_{\pi,\mu,\theta}(x) = \sum_{k=1}^K \pi_k \frac{1}{(2\pi\sigma_k^2)^{d/2}} \exp\Big(- \frac{1}{2\sigma_k^2} \|x - \mu_k\|^2 \Big).$$

    Represent graphically a typical such distribution for $d = 1$ and $K = 2$. Can such a distribution handle the shortcomings of K-means? What would be approximately good parameters for the data above?

32. (2 points) By applying Jensen's inequality, show that for any positive vector $a \in (\mathbb{R}_+^*)^K$, then

$$\log \sum_{k=1}^K a_k \geqslant \sum_{k=1}^K \tau_k \log \frac{a_k}{\tau_k}$$

    for any $\tau \in \Delta_K$ (the probability simplex), with equality if and only if $\tau_k = \frac{a_k}{\sum_{k'=1}^K a_{k'}}$.

33. (4 points) We assume that we have n independent and identically distributed observations x_i of X for $i = 1, \ldots, n$. Show that

$$\log p_{\pi,\mu,\theta}(x) = \sup_{\tau \in \Delta_K} \sum_{k=1}^{K} \tau_k \log \left[\pi_k \frac{1}{(2\pi\sigma_k^2)^{d/2}} \exp\Big(-\frac{1}{2\sigma_k^2} \|x - \mu_k\|^2 \Big) \right] - \sum_{k=1}^{K} \tau_k \log \tau_k.$$

Provide an expression of the maximizer τ as a function of π, μ, θ and x.

Provide a probabilistic interpretation of τ as a function of x.

34. (2 points) Write down a variational formulation of the log-likelihood ℓ of the data $(x_1, \ldots, x_n)$ in the form

$$\ell = \sum_{i=1}^{n} \sup_{\tau_i \in \Delta_K} H(\tau_i, x_i, \pi, \mu, \sigma)$$

for a certain H.

35. (4 points) Derive an alternating optimization algorithm for optimizing $\sum_{i=1}^{n} H(\tau_i, x_i, \pi, \mu, \sigma)$ with respect to τ and (π, μ, σ).

36. (1 point) What are its convergence properties?