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.

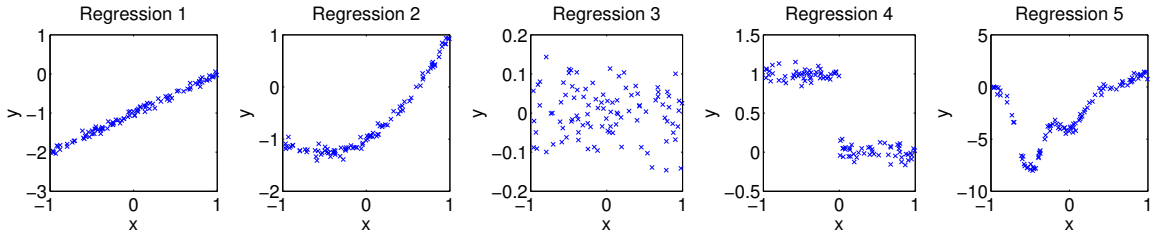
1 "Question de cours" (16 points)

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.

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 $\tilde{X} := \left[1, X\right]$ the $(n \times 2)$ matrix where the first column is $(1, \ldots, 1)^\top \in \mathbb{R}^n$, we have $\hat{\beta}_n = (\tilde{X}^\top \tilde{X})^{-1}\tilde{X}^\top Y$.

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?

Solution: Polynomial regression for one-dimensional inputs needs to compute $(Z^\top Z)^{-1}Z^\top X$ where $Z = [1, X, X^2, \ldots, X^d]$ 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. (2 points) 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. (2.5 points) 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. (1 point) 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. (1 point) Provide one solution to deal with over-fitting.

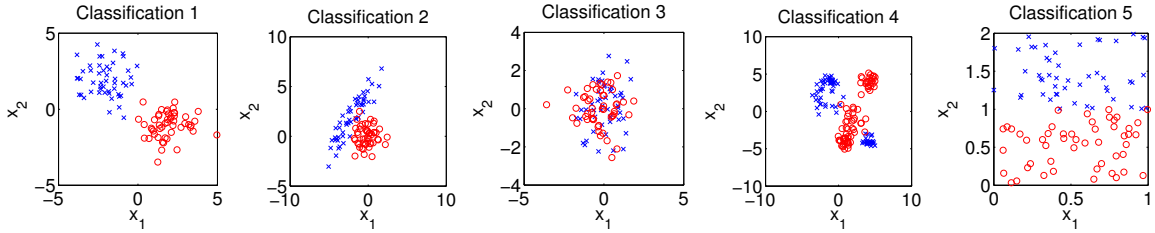
Solution: A solution is to use cross-validation to calibrate the hyper-parameters to regularize enough the methods (such as the regularization parameter λ in KRR or the bandwidth σ). Cross-validation can also be used to select the best model among (a-e).

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?

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

Solution: Linear discriminant analysis assumes that the data is generated from a mixture of Gaussian: given a group (label Y_i) the variables X_i are independently sampled from the same multivariate Gaussian distribution. Another common assumption that simplifies the computation of the solution is the homogeneity of variance between groups.

9. (2.5 points) 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

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.

Solution: Minimizing a strongly-convex function over a compact set always leads to a unique minimizer.

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

Solution: We have $x = z$ which is the feasible unconstrained minimizer.

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

Solution: If not, the points $y(\varepsilon) = x + \varepsilon(z - x)$ are such that $\|y(\varepsilon)\|_1 \leqslant \mu$ for all $\varepsilon > 0$ sufficiently small, and $\|y(\varepsilon) - z\|^2 < (1 - \varepsilon)\|x - z\|^2 + \varepsilon \times 0$, by Jensen's inequality. Then x cannot be the minimizer.

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:

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

Solution: If $z_i > 0$, and $x_i < 0$, then $|z_i - 0| < |z_i - x_i|$, and thus replacing x_i by 0 leads to a strictly better solution, hence $z_i > 0$ implies that $x_i \geqslant 0$. Similarly, $z_i < 0$ implies that $x_i \leqslant 0$. Also $z_i = 0$ implies that $x_i = 0$.

Therefore, if $\varepsilon \in \{-1, 1\}^n$ is a vector of signs of z, then, we can take $u = |z|$ (taken componentwise) and recover the solution x as $\varepsilon \circ u$.

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?

Solution: We can write the Lagrangian as

$$\mathcal{L}(y, \beta) = \frac{1}{2}\|y - u\|_2^2 + \beta\big(\sum_{i=1}^n y_i - 1\big).$$

Strong duality holds because the objective is convex, the contraints linear and there exists a feasible point (and Slater's conditions are satisfied).

Minimizing with respect to y leads to $(y_*)_i = (u - \beta_i)_+$ and the dual function

$$q(\beta) = -\frac{1}{2}\sum_{i=1}^n \max\{0, u_i - \beta\}^2 + \frac{1}{2}\|u\|_2^2 - \beta.$$

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 $O(n \log n)$.

Solution: The function $\max\{0, u_j - \beta\}^2$ is piecewise quadratic and continuously differentiable, with a break point of the derivative at $\beta = u_j$ for all j. Moreover, $q(\beta)$ tends to $-\infty$ when $\beta \to \infty$ and $\beta \to -\infty$, so there is a minimizer.

We have $q'(u_j) = -\sum_{i=1}^n 1_{u_j \leqslant u_i}(u_j - u_i) - 1$. If we assume that after we sort all u_i in time $O(n \log n)$ and reorder them so that $u_1 \geqslant u_2 \geqslant \cdots \geqslant u_n$, then

$$q'(u_j) = -\sum_{i=1}^{j-1}(u_j - u_i) - 1 = (j-1)u_j - 1 + \sum_{i=1}^{j-1} u_i.$$

We can then compute the non-creasing (by concavity) sequence of derivatives in $O(n)$ so find the interval where it is zero. All of this in $O(n)$.

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?

Solution: Assuming H to be symmetric, f is strictly convex if and only if H is positive definite (only strictly positive eigenvalues). The minimum is then unique equal to $\theta_* = H^{-1}g$.

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 $\theta_0 - \theta_*$?

Solution: We have $f'(\theta) = H\theta - q = H(\theta - \theta_*)$, and thus $\theta_t - \theta_* = (I - \gamma H)(\theta_{t-1} - \theta_*) = (I - \gamma H)^t(\theta_0 - \theta_*)$.

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

Solution: We have $f(\theta) - f(\theta_*) = \frac{1}{2}(\theta_0 - \theta_*)^\top H(\theta_0 - \theta_*)$.

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?

Solution: We have $f(\theta_t) - f(\theta_*) = \frac{1}{2}(\theta_0 - \theta_*)^\top H(I - \gamma H)^{2t}(\theta_0 - \theta_*)$, and the eigenvalues of $(I - \gamma H)^{2t}$ are between 0 and $(1 - \gamma\mu)^{2t}$. The best is $\gamma = 1/L$.

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?

Solution: We have $f(\theta_t) - f(\theta_*) = \frac{1}{2}(\theta_0 - \theta_*)^\top H(I - \gamma H)^{2t}(\theta_0 - \theta_*)$, and the eigenvalues of $H(I - \gamma H)^{2t}$ are between 0 and $\max_{\alpha \in [0,L]} \alpha \exp(-2\gamma\alpha t) = \frac{1}{2\gamma t}\max_{u \geqslant 0} ue^{-u} \leqslant \frac{1}{4\gamma t}$. The best is $\gamma = 1/L$.

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 $(\varepsilon_k)_{k \leqslant t}$?

Solution: We have

$$\theta_t - \theta_* = (I - \gamma H)(\theta_{t-1} - \theta_*) - \gamma \varepsilon_t = (I - \gamma H)^t (\theta_0 - \theta_*) - \gamma \sum_{k=1}^{t} (I - \gamma H)^{t-k} \varepsilon_k.$$

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

Solution: $\mathbb{E}\theta_t$ follows exactly the gradient descent recursion.

23. (3 points) Show that

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

Solution: This is simply computing the expectation and using the independence of the sequence (ε_k).

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 $1/\gamma$.

Solution: This is simply summing a geometric series and expanding $(I - \gamma H)^2$, and using that $2 - \gamma L \geqslant 1$.

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 $t > 0$,

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

Solution: This is simply the consequence of previous questions.

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.

Solution: This is simply the consequence of previous questions. $\gamma^* = \frac{\|\theta_0 - \theta_*\|}{2\sqrt{\operatorname{tr} C}}$, leading to the bound

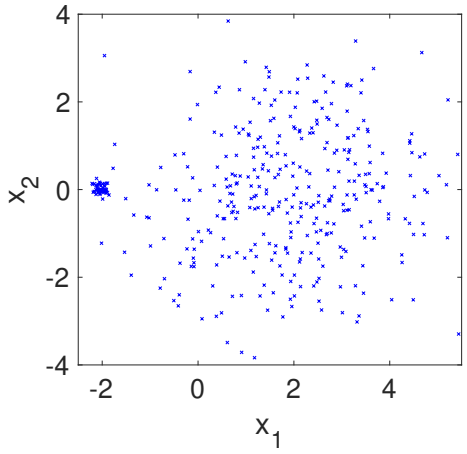
$$\frac{\|\theta_0 - \theta_*\|\sqrt{\operatorname{tr} C}}{2\sqrt{t}}.$$

The optimal step-size depends on potentially unknown quantities and the scaling is $O(1/\sqrt{t})$, which is worse than the deterministic case in $O(1/t)$.

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.



Solution: K-means would cut the large cluster in two.

We consider a probabilistic model on two variables X and Z, where $X \in \mathbb{R}^d$ and $Z \in \{1, \dots, K\}$. 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 $(x, z) \in \mathbb{R}^d \times \{1, \dots, K\}$.

Solution: We have: $\log p(x, z) = \log p(z) + \log p(x|z) = \log \pi_z - \log(2\pi\sigma_z^2)^{d/2} - \frac{1}{2\sigma_z^2}\|x - \mu_z\|^2$.

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

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.

Solution: We have, because of the i.i.d. assumption:

$$\begin{aligned}\ell &= \sum_{i=1}^{n}\log p(x_i, z_i)\\ &= \sum_{i=1}^{n}\left\{\log\pi_{z_i} - \log(2\pi\sigma_{z_i}^2)^{d/2} - \frac{1}{2\sigma_{z_i}^2}\|x_i - \mu_{z_i}\|^2\right\}\\ &= \sum_{k=1}^{K}\sum_{i=1}^{n}\delta(z_i = k)\left\{\log\pi_k - \log(2\pi\sigma_k^2)^{d/2} - \frac{1}{2\sigma_k^2}\|x_i - \mu_k\|^2\right\}.\end{aligned}$$

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

Solution: ML decouples. We have $\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n}\delta(z_i = k)$, which is the proportion of observed class k.
Moreover, $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}\delta(z_i = k)x_i$, which is the mean of the data points belonging to class k.
Finally, we get $\hat{\sigma}_k^2 = \frac{1}{dn}\sum_{i=1}^{n}\delta(z_i = k)\|x_i - \hat{\mu}_k\|^2$.

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?

Solution: We use: $p(x) = \sum_{k=1}^{K}p(x, k)$ to get to the expression. The data were generated with $\pi = (1/8, 7/8)$, $\mu_1 = (-2, 0)$, $\mu_2 = (+2, 0)$, $\sigma_1 = 1/10$, $\sigma_2 = 3/2$ and $n = 400$.

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 $\tau_k = \frac{a_k}{\sum_{k'=1}^{K}a_{k'}}$.

Solution: This is simply Jensen's inequality for the logarithm, with

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

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.

Solution: We have:

$$\begin{aligned} \log p(x) &= \log \sum_{k=1}^{K} \pi_k p(x|Z=k) \\ &= \log \sum_{k=1}^{K} \tau_k \frac{\pi_k p(x|Z=k)}{\tau_k} \\ &\geqslant \sum_{k=1}^{K} \tau_k \log \frac{\pi_k p(x|Z=k)}{\tau_k}. \end{aligned}$$

There is equality if and only if $\tau_k = \frac{\pi_k p(x|Z=k)}{\sum_{k'=1}^{K} \pi_{k'} p(x|Z=k')}$, which is exactly $\tau_k = p(Z=k|x)$ (which acts a soft-clustering of each input x, as oppose to K-means that performs hard clustering).

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.

Solution: This is applying the previous question for all i, with

$$H(\tau_i, x_i, \pi, \mu, \sigma) = \sum_{k=1}^{K} \tau_{ik} \log \left[\pi_k \frac{1}{(2\pi\sigma_k^2)^{d/2}} \exp\Big(- \frac{1}{2\sigma_k^2} \|x_i - \mu_k\|^2 \Big) \right] - \sum_{k=1}^{K} \tau_{ik} \log \tau_{ik}.$$

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).

Solution: The maximization with respect to τ (often called the "E-step") leads to $\tau_{ik} = p(Z = k|x_i)$ for the current value of the parameters (π, μ, σ).

The maximization with respect to (π, μ, σ) (often called the "M-step") leads to $\hat{\pi}_k = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}$.

Moreover, $\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik} x_i$.

Finally, we get $\hat{\sigma}_k^2 = \frac{1}{dn} \sum_{i=1}^{n} \tau_{ik} \|x_i - \hat{\mu}_k\|^2$.

The M-step is simply the same as estimating parameters for full observations with $\delta(z_i = k)$ replaced τ_{ik}.

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

Solution: This is an ascent algorithm. No guarantees of convergence to any global maximum of the log likelihood. It is often called the "Expectation-Maximization" algorithm.