Exam of
Foundational Principles of Machine Learning (FPML)
December 16, 2022 – 3h

DON'T RETURN THIS SHEET BEFORE YOU ARE ALLOWED TO DO SO!
(you can write your name on the blank papers in the meantime)

General advice:

- **Authorized documents: 6 pages of personal notes.**
- Do not hesitate to do the exercises in any order you like: start with the ones you feel are quick to deal with.
- When you are allowed to start, before you start the first exercise, go through the subject quickly. In each exercise, the most difficult question is not necessarily the last, please feel free to skip some questions. Don't hesitate to go and scrap off points where they are easy to take. (vous pouvez "aller grapiller les points")
- The grading points (scale) is indicative, if the exam is too long a correction factor will be applied. So don't panic in front of the length, what you do, do it right! Also, you may notice that points sum up to 21 (5+5.5+7.5+4) instead of 20, so, I will do *something*.
- **French:** Vous êtes autorisés à composer en Français. (Y compris en insérant des mots techniques comme overfitting ou regularization en anglais quand vous ne savez pas la traduction).
- **French:** si certains bouts de l'énoncé ne sont pas clairs, je peux les traduire ! N'hésitez pas à demander si vous n'êtes pas sûrs.
- Calculators not allowed (and useless). No electronic device allowed (cell phone, etc).
- At the end, we will collect your papers. You can leave after you have returned your paper.

DON'T RETURN THIS SHEET BEFORE YOU ARE ALLOWED TO DO SO!
(you can write your name on the copies in the meantime)

1 Lecture related and independent questions (5 points)

1. (0.25 pt) Does PCA takes into account the labels of data points?
2. (0.25 pt) What does SVM stand for? (2 answers accepted)
3. (0.5 pt) Let's assume we have data and a model we trained. How can we estimate whether more data would be helpful, when we do not yet have this additional data?
4. (0.5 pt) In PCA, what is maximized, or minimized? (2 answers accepted). Be precise in your answer with words, or, write the answer mathematically.
5. (0.5 pt) What are the benefits or uses of cross-validation? (give two)
6. (1 pt) When you have overfitting, what are the things you can do? (assuming we keep the same family of models). Cite as many possible solutions as you know, and each time, quickly explain your choice (1-2 lines per "solution" to overfitting).
7. (1 pt) Explain the idea of the SVM. Introduce the proper key concepts and explain what the SVM models maximize.
8. (0.5 pt) In ML, many learning algorithms consist essentially in Gradient Descent, a rather simple minimization algorithm. Explain with words why this is not such a bad idea, referring to the idea of train and test sets (we can omit hyper-parameters and the validation set in this explanation, for simplicity).
9. (0.5 pt) Formalize the previous answer in terms of argmin, Losses, GD, and any notation you find useful.

2 Maximum A Posteriori (MAP) (5.5 points)

We want to compute the MAP estimates of the parameters μ, σ for a random variable X that we model as following a Gaussian law, $\rho(x) \sim \mathcal{N}(\mu, \sigma)$. We recall that $\mathcal{N}(\mu, \sigma) : \rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.

We have access to a data set $\widetilde{X} = \{x_1, \dots, x_N\}$ of empirical observations. We may refer to the empirical mean as $\overline{x} = \frac{1}{N}\sum_n^N x_n$ and to the empirical variance as $\overline{V}[\widetilde{X}] = \overline{\sigma}^2 = \frac{1}{N}\sum_n (x_n - \overline{x})^2$.

The event "I observe the data is $\widetilde{X}$" can be written $X = \widetilde{X}$.

1. (2 pt) At first we introduce a Gaussian prior for the parameter μ: $\rho(\mu) = \mathcal{N}(0, \tau)$. Do compute the MAP estimate μ_{MAP} from the start, i.e. from the definition, recalling the fundamental steps (that are general to all cases) as well as the computations that apply to this precise case.
2. (0.25 pt) Without much computation, say what is σ_{MAP} (you can explain well or draft a few lines of computation)
3. (0.25 pt) Interpret (comment on) the limit $N \to \infty$.
4. (0.25 pt) Interpret (comment on) the limit $\tau \to 0$.
5. (0.25 pt) Interpret (comment on) the limit $\tau \to \infty$.
6. (2 pt) Now we forget this prior on μ and only assume a prior on σ. Precisely, we use an exponential prior for σ (which by definition, is positive), σ: $\rho(\sigma) = \lambda e^{-\lambda\sigma}$, where $\lambda > 0$. Remember that $\mathbb{E}[\rho_\lambda] = \int_0^\infty \sigma\ \lambda e^{-\lambda\sigma} \mathrm{d}\sigma = 1/\lambda$. Compute the MAP estimate σ_{MAP}. Computation is slightly heavier but conceptually not more difficult than in the first question.
7. (0.5 pt) Now, we assume both priors simultaneously, i.e. we have the joint prior distribution: $\rho(\mu, \sigma) = \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(\mu)^2}{2\tau^2}} \lambda e^{-\lambda\sigma}$.
   Does μ_{MAP} change at all, and if so, how ?
   Does σ_{MAP} change at all, and if so, how ?
   (Note: I do not expect a lot of computation in this last question, just a bit of reasoning.)

3 Binary classification with a few neurons (7.5 pts)

We have a dataset $X, Y = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_N, y_N)\}$ where Y encodes the binary labels. The data is D dimensional, $\vec{x} \in \mathbb{R}^D$. To classify the data, we use the following model and Loss:

$$\hat{y}(\vec{x}_n) = f_\theta(\vec{x}_n) = \tanh(v\ ReLU(\vec{w} \cdot \vec{x}_n + a) + b) \tag{1}$$

$$\mathcal{L} = \frac{1}{N} \sum_n^N (\hat{y}(\vec{x}_n) - y_n)^2 \tag{2}$$

where ReLu is the so-called *Rectified Liner Unit*, i.e. $ReLU(z) = max(0, z)$. We may introduce the convenient notations:

$$f_\theta^{(1)}(\vec{x}_n) = \vec{w} \cdot \vec{x}_n + a \tag{3}$$

$$f_\theta^{(2)}(\vec{x}_n) = v\ ReLU(f_\theta^{(1)}(\vec{x}_n)) + b = v\ ReLU(\vec{w} \cdot \vec{x}_n + a) + b \tag{4}$$

$$\text{so that we have: } f_\theta(\vec{x}_n) = \tanh(f_\theta^{(2)}(\vec{x}_n)) \tag{5}$$

We recall that the hyperbolic tangent function is defined by: $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$. It goes from -1 at $z \sim -\infty$ to $+1$ at $z \sim +\infty$. Its derivative is (for a generic smooth function u): $\frac{\partial}{\partial z} \tanh(u(z)) = \frac{4u'(z)}{(\cosh(u(z)))^2}$, where $\cosh(z)$ is the hyperbolic cosine, $\cosh(z) = \frac{e^z + e^{-z}}{2}$. We know that $\cosh(z) \geq 1, \forall z$.

We denote $H(z)$ the Heaviside function (step function) : $H(z) = \begin{cases} 1 \text{ if } z \geq 0 \\ 0 \text{ if } z < 0 \end{cases}$.

You can introduce other notations if you find them useful.

1. (0.5 pt) This question is independent from the next ones. What is the maximal Loss that a single example can contribute to the Loss? What is its minimal value? Considering this observation, do you think this model is very sensitive to outliers?
2. (0.25 pt) List the parameters θ of this model, and for each, say what space it is in ($\mathbb{R}, \mathbb{R}^D$, etc).
3. (0.25 pt) Compute the derivative $\frac{\partial}{\partial z} ReLU(u(z))$, for a generic function $u(z)$ (assumed derivable).
4. (0.75 pt) Compute $\nabla_b \mathcal{L}$. Ideally, try to derive the steps in a generic way as far as you can in the computation (this is to help you for the next questions).
5. (0.75 pt) Compute $\nabla_v \mathcal{L}$ (re-use previous question's result).
6. (0.75 pt) Compute $\nabla_a \mathcal{L}$ (re-use previous question's result).
7. (0.75 pt) Compute $\vec{\nabla}_{\vec{w}} \mathcal{L}$ (re-use previous question's result).
8. (0.5 pt) Write down $\vec{\nabla}_{\vec{w}} \mathcal{L}$ in terms of $\nabla_b \mathcal{L}$ (making it appear explicitly).
9. (0.25 pt) What should be our choice of encoding of the binary labels y_n? In other words, which are the 2 values the y_n's should take?
10. (0.25 pt) What should be the decision function (readout), i.e. what is the expression of y_{predict} as a function of $\hat{y}$?
11. (0.5 pt) Compute $\nabla_b \Big(\frac{1}{N} \sum_n^N (y_{\text{predict}}(\vec{x}_n) - y_n)^2 \Big)$? What is the problem with this loss ?
12. Priors. We now have the idea that the data comes from the following process:

$$y_n = \tanh(v\ ReLU(\vec{w} \cdot \vec{x}_n + a) + b) + \varepsilon_n \tag{6}$$

where ε_n are noise terms, each ε_n is an i.i.d. Gaussian variable: $\rho(\varepsilon) = \mathcal{N}(0, \sigma)$, i.e. $\rho(\varepsilon) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}}$. Additionally, we have a Gaussian prior on v: $\rho(v) = \mathcal{N}(0, \lambda)$.

   (a) (0.5 pt) Can you guess the result of a MAP estimate for v ?

   (b) (1 pt) Prove your guess, i.e. perform the MAP reasonning, being as rigorous as possible.
   Note: you cannot and are not asked to compute v_{MAP} explicitly, you should just write the problem that needs to be solved numerically to get v_{MAP}, and notice to what it corresponds to.

   (c) (0.5) Guess again what happens if we have a Gaussian prior on each component w_d of the vector $\vec{w}$.

   Note: in deep neural networks, typically, people use such priors.

4 Lasso Regularization (simple) (4 pts)

The first two questions are not about Lasso, but are here as preliminary, to help.

1. (0.5) Solve the minimization problem: $\text{argmin}_w(\frac{1}{2}(w-a)^2 + \lambda w^2)$, where $w \in \mathbb{R}, a \in \mathbb{R}$

2. (0.5) Solve the minimization problem: $\text{argmin}_{\vec{w}}(\frac{1}{2}||\vec{w}-\vec{a}||^2 + \lambda||\vec{w}||^2)$, where $\vec{w} \in \mathbb{R}^D, \vec{a} \in \mathbb{R}^D$. Take the time to first find the solution for the component w_1, explicitating the scalar products (or norms) as sums, and only then, generalize.

3. (2 pts) Solve the minimization problem: $\text{argmin}_w(\frac{1}{2}(w-a)^2 + \lambda|w|)$, where $w \in \mathbb{R}, a \in \mathbb{R}$, and $|.|$ is the absolute value.
   You should assume that your pseudo-gradient $\frac{\partial}{\partial w}|w|$ is in the range $[-1, 1]$, and check that you find a solution for any value of a.
   Hint: you should discuss the cases where the solution is $w \neq 0$ and where it is $w = 0$, separately.

4. (1 pt) Solve the minimization problem: $\text{argmin}_{\vec{w}}(\frac{1}{2}||\vec{w}-\vec{a}||^2 + \lambda|\vec{w}|_1)$, where $|.|_1$ is the $L1$ norm, i.e. $|\vec{w}|_1 = \sum_d^D |w_d|$.
   Hint: you can generalize from previous cases, not necessarily re-writing all explanations.