EXAMEN FONDAMENTAUX DU MACHINE LEARNING
MASTER IASD - 2022

Duration: 3 hours. You are allowed to bring 4 sheets of paper recto/verso (manuscripted or not).

1. Exercice - SVMs (3 points)

Consider a set of labeled data $(\mathbf{x}_i, y_i)_{i=1}^n$ and $y_i \in \{+1, -1\}$. We want to solve a variant of the SVM problem. The underlying optimization problem is:

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \xi_i^2$$
$$s.t. \quad y_i \left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1 - \xi_i \quad \forall i \in \{1 \dots N\}$$

(1) Write the lagrangian of this problem and the optimality conditions of the lagrangian with respect to the variables $\mathbf{w}$ and b.
(2) Give the condition under which an example is a support vector
(3) Write the dual of this problem.

2. Exercice - Scoring Losses (5 points)

In this exercise, we define the label space as $\mathcal{Y} = \{-1, 1\}$ and the prediction space $\hat{\mathcal{Y}} = \mathbb{R}$. Let us define the squared hinge loss $\ell(\hat{y}, y) = \max(0, 1 - \hat{y}y)^2$.

For your information, the derivative of this loss is

$$\frac{d\ell(\hat{y}, y)}{d\hat{y}} = \begin{cases} 2(\hat{y} - y) & \text{if } y\hat{y} < 1 \\ 0 & \text{otherwise} \end{cases}$$

(1) Is this loss a phi-loss ? Is this loss calibrated ? why ?
(2) What is the Bayes predictor for this loss ? what is its Bayes risk ?
(3) Show how to derive a criterion for decision trees based on this loss.
(4) Show how to apply gradient boosting on this loss (give the complete algorithm, with sufficient details to implement it in Python)
(5) **Class Probability Estimate (CPE) functions.** Let h be scoring classifier $h : \mathbb{R}^d \mapsto \hat{\mathcal{Y}}$. Let σ be the sigmoid function. Let $g = \sigma \circ h$, the CPE classifier obtained by putting a sigmoid function on top of h. Show that learning the scoring classifier h with the scoring loss ℓ is equivalent to learning the CPE function g with a CPE-loss ℓ^{CPE}. What is this CPE-loss ? (bonus question: show ℓ^{CPE} is not proper)

3. Exercice - Chernoff bound (3 points)

(1) Consider a biased coin with the probability of getting heads being an unknown parameter p, which is known to be at least a, for some $a > 0$. A natural procedure for estimating the coin bias is to flip the coin N times, and estimate p as the fraction of times it lands on head. Denote this estimate by $\hat{p}$ and suppose that for a given parameter we want to have $|p - \hat{p}| \leq \epsilon p$ with probability greater than $1 - \delta$.
How many flips do we need in function of a, δ and ϵ ?
(2) Show that with the same number of flips, we will have $|p - \hat{p}| \leq \epsilon$ with probability greater than $1 - \delta$.
(3) Assume we have a finite class of functions $\mathcal{F}$, on which we want to learn with the ERM principle using the 0/1 loss. Assume we know that *all* functions in $\mathcal{F}$ have a risk bigger than a. With the help of the union bound, use the above answer to show immediately if N is bigger than some formula you have to determine, then $\sup_{f\in\mathcal{F}} \left|R(f) - \hat{R}(f)\right| \leq \epsilon$ with probability at least $1 - \delta$.

4. Exercise - Rademacher/Pac/Mdl (3 points)

In this exercise, we will focus on the zero-one loss $\ell(\hat{y}, y) = \mathbf{1}[\hat{y} \neq y]$. Let $\mathcal{X} = \mathbb{R}$ be the space of examples and $\mathcal{Y} = \hat{\mathcal{Y}} = \{0, 1\}$ the space of predictions and labels. Let $\mathcal{P}oly$ be the class of polynomials mapping $\mathbb{R}$ to $\mathbb{R}$ and of arbitrary degree. For example, the polynomial $x \rightarrow x^{18} - 3x + 4$ belongs to this class, for $x \in \mathbb{R}$. Define the following class of functions $\mathcal{F} = \{\mathbf{1}[p(x) \geq 0] : p \in \mathcal{P}oly\}$ where $\mathbf{1}[\cdot]$ is the indicator function. Let $\mathcal{G} = \ell \circ \mathcal{F} = \{(x, y) \rightarrow \ell(f(x), y) : f \in \mathcal{F}\}$. Consider a dataset $S = (x_i, y_i)_{i=1}^{N}$ drawn i.i.d. from an unknown distribution. Recall that if $x_i \neq x_j$ for all $i \neq j$, there exists a polynomial of degree $N-1$ going exactly through all the points in S.

(1) **Rademacher Complexity.** Show (with as much details as needed) what is the Rademacher complexity of $\mathcal{F}$ on S. Deduce the Rademacher complexity of $\mathcal{G}$. What can we deduce from this result if we wanted to implement a learning algorithm for the class $\mathcal{F}$ based on ERM ?

(2) **Minimum Description Length Principle.** The Chernoff-Hoeffding says that for a *single* given function f, if the loss function is the 0/1 loss, then we have $\left|R(f) - \hat{R}(f)\right| \leq \sqrt{\frac{1}{2N} \ln \frac{2}{\delta}}$ with probability at least $1 - \delta$, where R and $\hat{R}$ denote the risk and the empirical risk over the dataset S.
Define $\delta_k = \frac{1}{k!}$ for all $k \in \{1, 2 \ldots\}$. Note that $\sum_{k=1}^{\infty} \delta_k = 1 - e$ and that $\ln(k!) \simeq k \ln k - k$.
Show precisely what bound the implementation of the MDL principle on the class $\mathcal{F}$ with the 0/1 loss would give us. Explain in one short sentence the impact on learning of using this $\delta_k = \frac{1}{k!}$ rather than $\delta_k = 2^{-k}$.

5. Exercise - Latent models and Probabilistic PCA (2 points)

This exercise studies a simplified version of Probabilistic Principle Component Analysis, so we are in an unsupervised setting. Let $\mathcal{X} = \mathbb{R}^d$ be the space of examples. Let k be an integer lower than d. In the following, Id_k refers to the identity matrix of dimension $k \times k$. Assume that we have a dataset $(x_i)_{i=1}^{N}$, which has been generated according to the following generative model:

- For each $i \in \{1 \ldots N\}$
  (1) draw $z_i \sim \mathcal{N}(0, Id_k)$
  (2) draw $\epsilon_i \sim \mathcal{N}(0, Id_d)$
  (3) $x_i = W z_i + \epsilon_i$

Here, $W \in \mathbb{R}^{d \times k}$ is an unknown matrix, and $z_i \in \mathbb{R}^k$ is the unknown latent variable associated to example x_i.

Recall that the density of a multivariate normal law is:

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^{\top} \Sigma^{-1} (x - \mu)\right)$$

so in the specific case of Identity covariance matrix, we get

$$\mathcal{N}(x \mid \mu, Id_d) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}\|x - \mu\|^2\right)$$

(1) Write the complete joint density $p(x_i, z_i \mid W)$.

(2) We want to solve this problem with the variational-EM framework.
  (a) which kind of distribution $q_x(z)$ should you choose ?
  (b) Write the ELBO (Evidence Lower Bound).
  Here, I do not expect fully developped/simplified formulas from you. This would require lots of tedious calculations. I just want to see that you understand the ideas of variational EM, not that you master the calculus of gaussian densities, so just give a precise formula applied to this setting without developping everything.
  (c) Write the algorithm.