MACHINE LEARNING

EXAM TSIA-SD210 2022 (DURATION : 2H)

The only authorized document is a A4 sheet of paper with personal notes ; Pleas keep your answers precise and short.

1 - SUPERVISED CLASSIFICATION

We consider the probabilistic and statistical framework of supervised classification where X is a random vector on $\mathbb{R}^d$, $d \geq 1$ and Y is a binary random variable with values in $\{-1, +1\}$. A random sample $\mathcal{S}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ contains n independent copies of the pair (X, Y) of joint probability distribution P.

1. Define a classifier and recall the definition of risk. Define the problem of supervised binary classification (in ideal conditions) using the definition of risk.
2. Define the empirical risk of a classifier calculated using $\mathcal{S}_n$. Explain the principle of empirical risk minimization.
3. What is overfitting ? What is the underlying idea of methods designed to avoid it ?

2 - SUPPORT VECTOR MACHINES

We consider the framework of binary supervised classification.

1. What optimization problem do we need to solve in the primal space to find the optimal margin hyperplane, e.g. a linear SVM, when data are noisy ?
2. Write the dual formulation of this problem.
3. Give the definition of a positive definite kernel, and give an example of kernel.
4. Give the decision function computed by a SVM based on a kernel.

3 - ENSEMBLE METHODS

1. What is the bias-variance decomposition ? Explain the 3 terms. How is it useful for analyzing bagging ?
2. Give the pseudocode of the random forest ; specify and justify
   — a halting condition for growing each tree,
   — the number of trees.
   How will the bias/variance improve with regard to a single tree ? When and how does it improve over bagging ?

4 - INTRODUCTION TO DEEP LEARNING

1. Recall the update rule used in the gradient descent method when optimizing a loss function l which depends on parameters θ, using a learning rate η.

We now work with a one-hidden-layer neural network, taking as inputs data points $\mathbf{x} \in \mathbb{R}^p$, and outputting a scalar value $o \in \mathbb{R}$. The model is comprised of parameters $\{\mathbf{W}^h, \mathbf{w}^o\}$ where $\mathbf{W}^h \in \mathbb{R}^{p \times d}$ and $\mathbf{w}^o \in \mathbb{R}^d$ (we assume no bias for simplicity). The output is obtained as follows :

$$\mathbf{h} = \sigma(\mathbf{W}^{h^\top}\mathbf{x}) \quad \text{and} \quad o = \mathbf{w}^{o^\top}\mathbf{h}$$

where $\mathbf{h} \in \mathbb{R}^d$ and σ is the sigmoid activation function. We want to train this model on a regression task, using a dataset $(\mathbf{x}^{(i)}, y^{(i)})_{i=1}^n$ with $\mathbf{x}^{(i)} \in \mathbb{R}^p$ and $y^{(i)} \in \mathbb{R}$, $\forall i \in [|1, n|]$. Hence, we use the MSE loss function ; when computed on one training pair, it is expressed as :

$$l_{MSE}(\mathbf{x}^{(i)}, y^{(i)}) = (y^{(i)} - o^{(i)})^2$$

2. Using backpropagation, compute the gradient updates corresponding to one training pair $(\mathbf{x}^{(i)}, y^{(i)})$ for :
   — the components w_k^o of $\mathbf{w}^o$, $\forall k \in [|1, d|]$,
   — the components W_{jk}^h of $\mathbf{W}^h$, $\forall (j, k) \in [|1, p|] \times [|1, d|]$.
3. With gradient updates computed via backpropagation, optimization of deep neural network remains difficult. Give and explain the idea behind 3 innovations made to ease training of deep neural models.