Machine Learning Fundamentals
Final Exam
2018-2019

Massih-Reza Amini

Duration: 2 hours, authorised documents: Slides of the course

1. (1 pt) Describe the *Structural Risk Minimization* principle.

2. (1 pt) Why a small training error is not sufficient to determine the efficiency of a learning algorithm?

3. (2 pt) Consider the following training set

$$S = \{((+1, +1), +1), ((-1, +1), +1), ((-1, -1), -1), ((+1, -1), +1)\}$$

where for each pair (for ex. $((+1, +1), +1)$), the first element corresponds to the vector representation of an observation (here $(+1, +1)$) and the second term to its class (here $+1$). We apply the perceptron algorithm to separate the observations of both classes, by supposing that the initial weights $\omega^{(\mathbf{0})} = (0, 0, 0)$; and the learning rate $\eta = 1$. Turn the algorithm on this example by showing the weights obtained after each update.

4. (2 pt) We apply the Adaboost algorithm over a training set of size 10;

$$S = \{(\mathbf{x}_i, y_i); i \in \{1, \ldots, 10\}\} \in (\mathcal{X} \times \{-1, +1\})^{10}$$

   4.1 At step 1 examples are assigned a uniform weight: $\forall i, D_1(i) = \frac{1}{10}$. We suppose that after the training phase, the first classifier $h_1 : \mathcal{X} \to \{-1, +1\}$ misclassifies 3 examples of S.
   Estime the error $\epsilon_1 = \sum_{i:h_1(\mathbf{x}_i)\neq y_i} D_1(i)$ and deduce the weights α_1 associated to h_1 found by the algorithme.

   4.2 Estimate new weights D_2 for the misclassified and well classified examples by h_1.

5. (12 pt) We consider a mono-label multi-class classification problem where observations $\mathbf{x} = (n_1, n_2, \ldots, n_d) \in \mathbb{N}^d$ are described by a discret vector of size d in which each characteristic is an integer. This corresponds for example to the representation of douments in the basis of number of times each word of

a given vocabulary occurs in the document or the representation of images in the basis of the intensity of their pixels.

Here, we suppose that oservations are generated by a probabilistic model as follows :each characterisitc $n_j, j \in \{1, \ldots, d\}$ of an observation $\mathbf{x}$ belonging to class $y = k$ is the realisation of a corresponding random variable X_j which has a probability of occurence equal to $\theta_{j|k}$

5.1 For a given observation $\mathbf{x} = (n_1, n_2, \ldots, n_d) \in \mathbb{N}^d$ and for the sake of presentation we note

$$\mathbb{P}(\mathbf{x} \mid y = k) = \mathbb{P}((X_1 = n_1, \ldots, X_d = n_d) \mid y = k)$$

In this case, show that $\forall k \in \{1, \ldots, K\}$,

$$\mathbb{P}(\mathbf{x} \mid y = k) = \mathbb{P}(X_1 = n_1 \mid y = k) \prod_{j=2}^{d} \mathbb{P}(X_j = n_j \mid X_1 = n_1, \ldots, X_{j-1} = n_{j-1}, y = k) \tag{1}$$

5.2 Explain why

$$\forall k \in \{1, \ldots, K\}, \mathbb{P}(X_1 = n_1 \mid y = k) = \binom{n}{n_1} \theta_{1|k}^{n_1} (1 - \theta_{1|k})^{n - n_1},$$

where $\binom{n}{n_1} = \frac{n!}{n_1!(n-n_1)!}$ is the binomial coefficient;
and $n = n_1 + n_2 + \ldots + n_d$.

5.3 From the two previous question deduce that

$$\mathbb{P}(\mathbf{x} \mid y = k) = \frac{n!}{n_1! n_2! \ldots n_d!} \prod_{j=1}^{d} \theta_{j|k}^{n_j}$$