Machine Learning Fundamentals
Final Exam
2019-2020

Massih-Reza Amini

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

Question 1 (4 pt)

1.1 Explain the principle of the Structural Risk Minimization.

1.2 Explain the Gradient Descent algorithm. In which case, the algorithm is ensured to converge?

1.3 What is the distance of $\mathbf{x} = (1, 1, 0)$ to the hyperplan of equation,

$$x_1 + x_2 + x_3 + 1 = 0,$$

in dimension 3?

1.4 Suppose that the empirical error of a prediction function f on a test set of size 1000 is $\hat{\mathcal{L}}(f, T) = 0.12$. What is the upper bound of the generalization error of f that holds with probability 0.99?

Question 2 (4 pt)
Consider the following binary classification and the training set of size 4 :

$$S = \left\{ \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, +1 \right); \left(\begin{pmatrix} -1 \\ 1 \end{pmatrix}, +1 \right); \left(\begin{pmatrix} -1 \\ -1 \end{pmatrix}, -1 \right); \left(\begin{pmatrix} 1 \\ -1 \end{pmatrix}, +1 \right) \right\}$$

2.1 Draw the points in a orthonormal basis of dimension 2.

2.2 We consider the perceptron algorithm and suppose that its initial weights and the bias (w_0) are null. Further suppose that the learning rate is fixed to 1. In this case what are the weights found by the algorithm after 4 updates, if the order of points that are taken is anti-clockwise when beginning from the point $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, i.e. $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ then $\begin{pmatrix} -1 \\ +1 \end{pmatrix}$ then $\begin{pmatrix} -1 \\ -1 \end{pmatrix}$ ... ?

2.3 What is the equation of the decision boundary?

2.4 Deduce the value of the margin.

2.5 The result of the Novikoff theorem is it respected here?

Question 3 (4 pt)
We consider an input space of dimension d, $\mathcal{X} \subseteq \mathbb{R}^d$. Estimate the gradients of the following surrogate losses with respect to the weights $\mathbf{w} \in \mathbb{R}^d$ of a linear prediction function $h_{\mathbf{w}} : \mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle$ for an example $(\mathbf{x}, y)$

$$\begin{aligned}
\ell_q(\mathbf{x}, y, \mathbf{w}) &= (y - \langle \mathbf{w}, \mathbf{x} \rangle)^2 \\
\ell_l(\mathbf{x}, y, \mathbf{w}) &= \ln(1 + e^{-y\langle \mathbf{w}, \mathbf{x} \rangle}) \\
\ell_e(\mathbf{x}, y, \mathbf{w}) &= e^{-y\langle \mathbf{w}, \mathbf{x} \rangle} \\
\ell_h(\mathbf{x}, y, \mathbf{w}) &= \max(0, 1 - y\langle \mathbf{w}, \mathbf{x} \rangle)
\end{aligned}$$

Question 4 (8 pt)

4.1 Explain the Adaboost algorithm seen in the course.

4.2 What is the role of the distribution D_t ?

4.3 After T rounds, the algorithm will learn T weak-classifiers $(f_t)_{1 \leq t \leq T}$, with their associated weights $(\alpha_t)_{1 \leq t \leq T}$ where the output of each weak classifier is binary in the set $\{-1, +1\}$.

4.4 How is obtained the final classifier F?

4.5 Explain why the empirical error of the final classifier F on a training set of size m; $S = \{(\mathbf{x}_i, y_i) \mid i \in \{1, \ldots, m\}\}$ is bounded by the following surrogate loss:

$$\mathcal{L}(F, S) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{y_i F(\mathbf{x}_i) \leq 0} \leq \frac{1}{m} \sum_{i=1}^{m} e^{-y_i \sum_{t=1}^{T} \alpha_t f_t(\mathbf{x}_i)}$$

where, $\mathbb{1}_\pi = 1$ if the predicate π is true; and 0 otherwise.

4.6 Show that

$$\frac{1}{m} \sum_{i=1}^{m} e^{-y_i F(\mathbf{x}_i)} = \sum_{i=1}^{m} Z_1 D_2(i) \prod_{t>1} e^{-y_i \alpha_t f_t(\mathbf{x}_i)} \tag{1}$$

$$\text{where, } \forall t, Z_t = \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y_i f_t(\mathbf{x}_i)} \tag{2}$$

4.7 By induction deduce that :

$$\frac{1}{m}\sum_{i=1}^{m} e^{-y_i F(\mathbf{x}_i)} = \prod_{t=1}^{T} Z_t$$

As the normalization terms are all positive, the minimization of the surrogate loss (Eq. 1) is then equivalent to the minimization of the normalization factors Z_t, at each iteration.

4.8 Considering the equation (Eq. 2); for which value of α_t – expressed with respect to the error $\epsilon_t = \sum_{i:y_i \neq f_t(\mathbf{x}_i)} D_t(i)$ – the factor Z_t is minimized?

4.9 For this particular value of α_t, what is the minimum value of Z_t?

4.10 Considering the following variable change $\gamma_t = \frac{1}{2} - \epsilon_t$; show that :

$$\forall t, Z_t = \sqrt{1 - 4\gamma_t^2}$$

4.11 For $\gamma_t < \frac{1}{2}$, we have $\sqrt{1 - 4\gamma_t^2} \leq e^{-2\gamma_t^2}$. In this case show that :

$$\frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{y_i \neq F(x_i)} \leq \prod_{t=1}^{T} Z_t \leq e^{-2\sum_{t=1}^{T}\gamma_t^2}$$

4.12 Explain why the misclassification error of the final classifier F is ensured to converge to 0 when the number of iterations T tends to infinity.