Ensemble Methods
Exam: October the 23rd

M2 Computer Science MALIA
2 hours

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Abstract

For this exam, you are allowed to use your handwritten personal notes only. Documents in any other form are not permitted, nor is electronic material.

Your telephone must be in silent mode and stored in your bag.

The exercises that make up this exam are independent and all answers to questions must be justified. The quality of your answers will be taken into account in the evaluation of your paper.

Exercise 1 : Study of a Regression Problem

Linear Regression

In this first part we consider the following penalized regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} \in \mathbb{R}^m, \mathbf{X} \in \mathscr{M}_{m,d+1}$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

We also consider the following dataset:

y	1	4	2	4.5	2.2	4.2
x_1	-2	3	-1	1	-4	3
x_2	0	2	0	0	0	-2

We also consider the following ridge regression optimization problem:

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^{d+1}} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2, \tag{1}$$

1. Show that this optimization problem is convex
2. Give the expression of parameter $\boldsymbol{\theta}$ according to $\mathbf{X}, \mathbf{y}$ and λ.
3. Give the values of the parameter $\boldsymbol{\theta}$ using the provided dataset in the following cases:
   (a) for any value $\lambda > 0$,
   (b) when $\lambda = 0$,
   (c) when $\lambda = 1$.

Quadratic regression

We now consider the following quadratic model

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \varepsilon,$$

where the parameters keep the same meaning as in the previous part. We do not focus on a regularized part.

Using the fact that

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 17.9 \\ 16.3 \\ 119.5 \end{pmatrix}$$

and

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.442 & -0.19 & -0.041 \\ -0.018 & 0.026 & 0.003 \\ -0.041 & 0.003 & 0.006 \end{pmatrix}.$$

4. Give the values of the parameter $\boldsymbol{\theta}$ (with a two order precision).

Prediction

In this question, we consider the following new example $\mathbf{x}_{new} \in \mathbb{R}^2$ defined by:

$$\mathbf{x}_{new} = \begin{pmatrix} 0.5 \\ -2 \end{pmatrix}.$$

5. Compute the prediction for the followings models:
   (a) the linear model when $\lambda = 0$.
   (b) the linear model when $\lambda = 1$.
   (c) the quadratic model.
   (d) a non-parametric estimator with the following kernel function

$$K_\eta(\mathbf{x}, \mathbf{x}') = \begin{cases} 3 & \text{if } \|\mathbf{x} - \mathbf{x}'\| \leq \eta = 3 \\ 0 & \text{otherwise} \end{cases}$$

   (e) Consider the ensemble hypothesis $H_4 = \frac{1}{4}\sum_{t=1}^{4} h_t$, where h_t refers to the four above mentioned model.
   Compute the predicted value of this ensemble hypothesis.

Exercise 2 : Bagging

1. Using your own words, explain the meaning of the following quantity and the importance it plays in the context of bagging. More precisely explain where this inequality comes from and what are the expectations of the algorithm used in the bagging procedure.

$$\frac{1}{T} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{X},y\sim\mathcal{Y}} \left[\sum_{t=1}^{T} (h_t(\mathbf{x}) - y)^2 \right] \geq \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{X},y\sim\mathcal{Y}} \left[(y - H_T(\mathbf{x}))^2 \right],$$

where $H_T = \sum_{t=1}^{T} h_t$, $y \in \mathbb{R}$ represents the value to predict, $\mathbf{x} \in \mathbb{R}^d$ is the feature vector and h_t are the base learners.

2. Explain the principle of double sampling mainly used in the Random Forest algorithm and which role it plays from both theoretical and practical point of view.

3. Recall how the validation procedure works with the Random Forest algorithm and, on average, the rate of examples that are used for the validation step.

Exercise 3 : Boosting

Preliminaries

1. Explain the difference between a strong and a weak classifier and provide an example for each of them.

2. In terms of Risk Minimization, what is the difference between the bagging and the boosting procedure?

Adaboost

We consider the Adaboost algorithm which aim to combine several classifiers in order to build a stronger one. The classifier we are going to dead with are linear ones. More precisely, we consider classifier h of the form

$$h(\mathbf{x}) = \text{sign}\left(\boldsymbol{\theta}^T\mathbf{x} + b\right),$$

where $\mathbf{x} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. We also consider the following dataset

y	-1	1	-1	-1	1	1	1	-1
x_1	-2	-1	-1	2	-4	-2	3	1
x_2	6	2	4	-3	-1	-2	-2	1

During the first iteration of our algorithm, the model parameters are given by $\boldsymbol{\theta} = (1, 1)$ and $b = -1$ and each example has a weight equal to $\frac{1}{m}$ where m denotes the number of examples.

3. Compute the error rate and the weight of this first learner.

4. Explain how the examples are reweighted in this algorithm.

Let us consider the following ensemble learner:

$$H(\mathbf{x}) = \text{sign}(\alpha_1 h_1(\mathbf{x}) + \alpha_2 h_2(\mathbf{x}) + \alpha_3 h_3(\mathbf{x})),$$

where $(\alpha_1, \alpha_2, \alpha_3) = (2, 3, 1.5)$ and $h_1(\mathbf{x}) = \text{sign}(3x_1 + 2), h_2(\mathbf{x}) = \text{sign}(2x_1 + 3x_2 - 4)$ and $h_3(\mathbf{x}) = \text{sign}(-x_1 + 2x_2 - 1)$.

Let $\mathbf{x}'_1$ and $\mathbf{x}'_2$ be two new instances defined by:

$$\mathbf{x}'_1 = (1, 0), \quad \text{et} \quad \mathbf{x}'_2 = (2, 3).$$

Predict the output of H_T of these two instances.

Gradient Boosting

In this section, we consider that $h : \mathcal{X} \to [-1, 1]$, *i.e.* its output is a real value this time.

We aim to solve the following minimization problem, *i.e.*, find the optimal weight of the base learner h_t

$$\alpha_t = \arg\min_{\alpha} \sum_{i=1}^{m} \exp[-y_i(H_{t-1}(\mathbf{x}_i) + \alpha h_t(\mathbf{x}_i))] = \arg\min_{\alpha} \sum_{i=1}^{m} w_i \exp[-y_i \alpha h_t(\mathbf{x}_i)], \quad (2)$$

where w_i is a pseudo-residual of our loss for the example $\mathbf{x}_i$.

5. Recall the definition of the pseudo residual for a given loss function ℓ.
6. Show that the solution of this minimization problem is given by

$$\alpha_t = \frac{1}{2} \ln \left(\frac{\sum_{i=1}^{m} (1 - y_i h_t(\mathbf{x}_i)) w_i}{\sum_{i=1}^{m} (1 + y_i h_t(\mathbf{x}_i)) w_i} \right).$$

Hint : you can use the fact that for all $i \in [\![1, m]\!]$

$$-y_i h_t(\mathbf{x}_i) = \frac{1 - y_i h_t(\mathbf{x}_i)}{2} - \frac{1 + y_i h_t(\mathbf{x}_i)}{2}$$