Ensemble Methods
Exam: October the 21st

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.

Exercise 1: 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.

4. Explain this remark using your knowledge of this algorithm and the relationships you've seen in class.

Exercise 2: 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	4	-3	-2	3	5	-1	6
x_2	6	-2	4	3	-1	-2	3	-1

During the first iteration of our algorithm, the model parameters are given by $\boldsymbol{\theta} = (1, 0)$ 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.
5. Let us consider the following weak learners:

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

   and $(\alpha_1, \alpha_2, \alpha_3) = (1, 2, 0.5)$.
   Let $\mathbf{x}'_1$ and $\mathbf{x}'_2$ be two new instances defined by:

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

   Predict the output of using the ensemble method generated by the Adaboost classifier.

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

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

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

6. Recall the definition of the pseudo residual for a given loss function ℓ.

7. 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 $i \in [\![1, m]\!]$

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

Extreme Gradient Boosting

The XGBoost algorithm is renowned for its extreme efficiency and speed. It is also of interest for its flexibility and potential adaptation to all loss functions.

1. On what principle does XGBoost operate with respect to the loss function used, *i.e.*, what assumptions are made about the loss function?
2. Write the problem when the considered loss function is the exponential one.
3. Based on the previous answer , and considering a dataset $S\{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$, find the general expression of value given by a leaf of the tree and apply it to the exponential loss function.

Exercise 3: F-measure

1. Recall the definition of precision, recall and F_β-measure. For the F-measure, give an expression as a function of the number P of positive data in the dataset, the number FN of false negatives and the number FP of false positives.
2. We will now denote $\mathbf{e} = (FN, FP)$ the error profile of a given classifier. We consider the following definition:

Définition: Pseudo Convexity

A real differentiable function f defined on an open convex set $\mathcal{C} \subset \mathbb{R}^q$ is said to be *pseudo-convex* if for every $\mathbf{e}, \mathbf{e}' \in \mathcal{C}$,

$$\langle \nabla f(\mathbf{e}), (\mathbf{e}' - \mathbf{e}) \rangle \geq 0 \implies f(\mathbf{e}') \geq f(\mathbf{e}),$$

where ∇f denotes the gradient of the function f.

Show that the F_β-measure is pseudo convex.