



Ensemble Methods
Exam: 9th October

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.

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

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 Algorithms

General questions on Boosting and Gradient Boosting

1. Explain in your own words the difference between *bagging* and *boosting*.
2. What is the main difference between Adaboost algorithm and a Gradient Boosting algorithm?
3. In gradient boosting, what is the role of the pseudo-residuals?

About XGBoost

We recall that XGBoost minimizes the following objective:

$$\mathcal{L}(t) = \sum_{i=1}^n \ell(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^t \Omega(f_k), \quad \text{with} \quad \Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_j w_j^2,$$

where ℓ is a loss function, and $\Omega(f)$ represents the regularization term.

4. Explain the meaning of each component in $\Omega(f)$.

5. How do the parameters γ and λ contribute to model regularization?
6. What would happen if both regularization terms were set to zero?
7. How does the inclusion of the Hessian (H_i) improve convergence compared to traditional gradient boosting?
8. XGBoost grows trees by optimizing a gain criterion. What does a positive gain signify?

Applications

Consider a node that can be split into two child nodes (Left and Right). The gradients (G) and Hessians (H) for the two nodes are:

Node	G	H
Left	-10	5
Right	-6	3

Regularization parameters: $\lambda = 1$, $\gamma = 2$.

The gain of a split is defined as:

$$Gain = \frac{1}{2} \left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} \right] - \gamma$$

9. Compute the gain by hand (show your steps).
10. Should the algorithm keep this split (recall that gain must be positive)?

In XGBoost, each leaf j receives an optimal weight:

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

Suppose a leaf has $G_j = -8$, $H_j = 4$, and $\lambda = 2$.

11. Compute the optimal leaf weight.
12. How would increasing λ affect the resulting weight and model variance?
13. Using the same data given at the beginning of this section.
 - (a) Recompute the gain when $\lambda = 0$ and $\gamma = 0$.
 - (b) Compare the two results (with and without regularization).
 - (c) Discuss qualitatively how regularization affects the decision to split and model complexity.

Exercise 3: PAC-Bounds

The aim of this exercise is to prove a General PAC Bayes-Bounds for real-valued losses.

To this aim, we introduce the following notations:

$$\mathbb{E}_D^\ell(h) = \mathbb{E}_{(x,y) \sim D} \ell(h(x), y) \text{ and } \mathbb{E}_S^\ell(h) = \frac{1}{m} \sum_{i=1}^m \ell(h(x_i), y_i),$$

where (x_i, y_i) denotes a couple of example and its associated binary label, ℓ is a loss function, D a distribution, h is a voter/hypothesis/classifier.

Preliminaries

1. Recall Markov's inequality and prove it.
2. Given a convex function f , recall Jensen's inequality.
3. Recall the definition of the KL-divergence.

Change of measure inequality For any set of voters $\mathcal{H}$, for any distribution P and Q on $\mathcal{H}$, and for any measurable function $\phi : \mathcal{H} \rightarrow \mathbb{R}$, show that we have

$$\mathbb{E}_{h \sim Q} [\phi(h)] \leq KL(Q \| P) + \ln \left(\mathbb{E}_{h \sim P} [e^{\phi(h)}] \right).$$

General PAC Bayes Theorem . The aim of this last part is to show the following result.

For any distribution D on $\mathcal{X} \times \mathcal{Y}$, for any set of voters $\mathcal{H}$, for any loss $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$, for any prior distribution P on $\mathcal{H}$, for any $\delta \in (0, 1]$, for any $m > 0$ (sample size) and for any convex function $f : [0, 1] \times [0, 1] \rightarrow \mathbb{R}$, we have

$$Pr_{S \sim D^m} \left(\forall Q \text{ on } \mathcal{H} : f \left(\mathbb{E}_{h \sim Q} \mathbb{E}_S^\ell(h), \mathbb{E}_{h \sim Q} \mathbb{E}_D^\ell(h) \right) \leq \frac{1}{m} \left[KL(Q \| P) + \ln \left(\frac{1}{\delta} \mathbb{E}_{S \sim D^m} \mathbb{E}_{h \sim P} e^{mf(\mathbb{E}_S^\ell(h), \mathbb{E}_D^\ell(h))} \right) \right] \right) \geq 1 - \delta$$

1. First, apply Markov inequality on the non negative random variable $e^{mf(\mathbb{E}_S^\ell(h), \mathbb{E}_D^\ell(h))}$, to introduce δ .
2. Apply the logarithm function and the change of measure result.
3. Use the Jensen's inequality to conclude.