Exercises: Ensemble learning

Gijs van Tulder / Jesse Krijthe

April 2021

Please upload the assignment in the form of a neatly formatted PDF ﬁle containing the names and student ids of the group members and your group’s answers. The grade for this assignment will be score+5 .

10

**Learning Objectives**

After this week’s lecture, reading and exercises, you will be able to

• explain Condorcet’s jury theorem

• explain the diﬀerent between trained and untrained combiners

• list several ways to combine the results of models and apply them.

• explain the diﬀerence between bootstrapping and random subspaces

• implement a random forest classiﬁer when given a decision tree implemen- tation

• eﬀectively apply a random forest classiﬁer to a dataset and interpret its parameters and results

• give an explanation why a random forest might perform better than a single decision tree

• give an informal deﬁnition of a weak learner

• list and explain the steps in the AdaBoost algorithm

• describe the diﬀerence between bagging and boosting

**Relevant Literature**

• Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). The Elements of

Statistical Learning (2nd ed.). Springer. Chapter 10 & 15

• Kuncheva, L. I. (2004). Combining Pattern Classifers. Methods and Al- gorithms. Wiley, Chichester. Chapter 4 & 5.

• Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). Foundations of

Machine Learning. The MIT Press. Chapter 6

• (Optional) Ho, T. K. (1998). The random subspace method for construct- ing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832-844.

• (Optional) Breiman, L. (2001). Random forests. Machine Learning, 45(1),

5-32.

**Exercises**

1. (15 points) Table 1 shows the posterior probability estimates, *pc*(*ω|x*), of three diﬀerent classiﬁers, for two diﬀerent classes *ω ∈ {*A*,* B*}*. Complete the table by ﬁlling in the values produced by the diﬀerent combiners and indicate which decision each combiner would make based on these values.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| *p*1(*ω|x*) | | *p*2(*ω|x*) | | *p*3(*ω|x*) | | Mean | | Max | | Min | | Prod | |
| A | B | A | B | A | B | A | B | A | B | A | B | A | B |
| 0.9 | 0.1 | 0.3 | 0.7 | 0.9 | 0.1 | 0.7 | 0.3 | 0.9 | 0.7 | 0.7 | 0.1 | 0.441 | 0.021 |
| 0.9 | 0.1 | 0.2 | 0.8 | 0.1 | 0.9 | 0.4 | 0.6 | 0.4 | 0.9 | 0.1 | 0.6 | 0.016 | 0.324 |
| 0.9 | 0.1 | 0.9 | 0.1 | 0.0 | 1.0 | 0.6 | 0.4 | 0.9 | 1 | 0 | 0.4 | 0 | 0.16 |
| 0.7 | 0.3 | 0.3 | 0.7 | 0.2 | 0.8 | 0.4 | 0.6 | 0.4 | 0.8 | 0.2 | 0.6 | 0.032 | 0.288 |
| 0.0 | 1.0 | 0.0 | 1.0 | 0.0 | 1.0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |

Table 1: Posterior predictions of three classiﬁers for a classiﬁcation problem with two classes (A, B), for 4 objects. Complete the table by calculating the values assigned by the diﬀerent combiners and determine what decision each combiner makes for each object.

2. (25 points) Suppose you are given an x-ray image of a patient, and you are given the task of choosing between the following three ways of making a diagnosis (yes vs. no bronchitis):

• An expert radiologist who has a probability of making a correct di- agnosis with probability *p* = 0*.*85.

• A group of 3 doctors, each of which has *p* = 0*.*75.

• A group of 31 medical students, each of which has *p* = 0*.*6. Please answer the following questions:

(a) (5 points) In the second case, what is the probability that all three doctors give the correct answer? What is the probability that at least 2 doctors make the right call? Combining these results, what is the probability that this group makes the right decision based on majority voting?

(b) (5 points) Come up with a general formula to calculate and/or code a simulation for the probability that *c* doctors with competence *p* make the correct decision by majority voting. Use it to calculate the probability of a correct decision for the group of medical students.

(c) (5 points) Make a graph of the probability of a correct decision for various sizes of the jury and diﬀerent competence levels (*p*) of the individual doctors.

(d) (5 points) Who has the highest chance to make the correct decision: the radiologist, the group of doctors or the group of students? How big does the group of medical students need to be to make the prob- ability of a correct decision (almost) equal to the prediction of the group of doctors?

(e) (5 points) If you did your computations correctly, you will have found that the probability of making a correct decision converges to 1 if the group of students is large enough. This is obviously unrealistic, but why? Explain your answer in terms of ensemble learning.

3. (20 points) Consider the following independent classiﬁers:

• A strong classiﬁer with the probability of making a correct decision with probability *p* = 0*.*75.

• An ensemble of 10 weak classiﬁers with probability *p* = 0*.*6.

(a) (5 points) Suppose that we combine the strong classiﬁer in an en- semble with the 10 weak classiﬁers. What is the probability of a correct decision in a majority vote if each classiﬁer’s vote has the same weight? Is the combined prediction better than that of the strong classiﬁer alone?

(b) (5 points) Instead of using an equal-weight majority vote, we can use a weighted majority vote in which the strong classiﬁer has a larger weight. Implement a function that computes, for a given weight *w* for the strong classiﬁer, the probability that the weighted majority vote results in the correct decision. Make a graph of the probability of a correct decision given diﬀerent weights. What is the optimal weight for the strong classiﬁer?

(c) (5 points) The AdaBoost.M1 algorithm provides a formula to com- pute the classiﬁer weights based on their error on the training set. Use the expected errors of the strong and weak classiﬁers to com- pute their respective weights. Compare the answer to the answer you found in the previous question.

(d) (5 points) Plot the weight given to a base-learner in the AdaBoost algorithm for diﬀerent values of the error the base-learner makes. Explain what you see. What does it mean for these weights if we assume the base-learners are weak-learners? What happens to the

weights if the probability of error of the base-learner is *>* 0*.*5 and why?

4. (15 points) In a few sentences, explain the diﬀerences between bootstrap- ping, random subspaces, and boosting.

5. (20 points) For this exercise you can choose to use random forests or AdaBoost. (If you are interested, feel free to compare both methods, but this is not required for the assignment.)

For your method of choice, ﬁnd out what widely used implementations are available in your favourite programming language and apply the method to a prediction problem you ﬁnd interesting (see, for instance the [UCI Machine Learning repository for](http://archive.ics.uci.edu/ml/index.php) interesting datasets). Write a short de- scription (min. 100 words) of your ﬁndings, including what dataset and implementation you used, how you set up your experiment, what the ef- fect of diﬀerent parameter settings was, what the performance was, which variables were important et cetera.

The cell indicated by starts have equal probabilities and B is selected based on other criteria like mean. It is figured out that the decision that made by Min and Max estimators are always equal. Besides, in this example output of Min and Max are the same as Prod.