# MTAT.03.227 Machine Learning
# Practice session 10 solutions

November 25-27, 2019

**Exercise 1. Boosting of decision stumps**

Answer the questions below for the dataset $D$ consisting of the following two-dimensional data points and binary labels:

$$x_1 = (1, 2), \quad x_2 = (0, 0), \quad x_3 = (-1, -1),$$
$$y_1 = +1, \quad y_2 = -1, \quad y_3 = +1$$

A decision tree with a single decision node is known as the *decision stump*. Let us define a learning algorithm $\mathcal{A}$ as a decision stump learner which selects its decision node as the one with minimal error. If several such exist then it chooses the one with the most positive predictions.

**(a)** Apply the AdaBoosting algorithm on the above data $D$ with $T = 3$ iterations and the decision stump learner $\mathcal{A}$. For this determine all the obtained instance weights $w_{ti}$ and model weights $\alpha_t$.

**(b)** Determine the final output score $H(x)$ of each of the instances. Are all instances classified correctly when thresholding $H(x)$ at 0? If not, how many iterations $T$ does it take to reach correct classifications?

**Solution.**

**(a)** As both of the features are monotonically decreasing along $x_1, x_2, x_3$, every possible decision stump outputs one of the following 6 sets of predictions for $x_1, x_2, x_3$: $+++$, $++-$, $+--$, $---$, $--+$, $-++$. These all have either 1 or 2 errors and among the predictions with 1 error $+++$ has the most positives. Therefore, in the first iteration of boosting the learning algorithm $\mathcal{A}$ chooses a decision stump which predicts all positives (that is, splitting based on whichever feature using any threshold, and predicting the positive class in both leaves). The error rate is $\epsilon_1 = 1/3$ and hence, the weight update factors are $1/(2\epsilon_1) = 3/2$ and $1/(2(1-\epsilon_1)) = 3/4$. The weights $(w_{21}, w_{22}, w_{23})$ for the second iteration are then equal to $(3/4 \cdot 1/3, 3/2 \cdot 1/3, 3/4 \cdot 1/3) = (1/4, 2/4, 1/4)$. Now predictions $+--$ and $--+$ both make a weighted error of $1/4$. We will assume that the learning algorithm chooses a decision stump predicting $+--$ (the results in case of choosing $--+$ are analogous).

Continuing these calculations we can find out that the instance weights $w_{ti}$, model predictions $h_t(x_i)$, weighted error rate $\epsilon_t$, weight update factors $1/(2\epsilon_t)$, $1/(2(1-\epsilon_t))$ and confidence $\alpha_t$ throughout the 3 steps are as follows:

| $t$ | $w_{t1}$ | $w_{t2}$ | $w_{t3}$ | $h_t(x_1)$ | $h_t(x_2)$ | $h_t(x_3)$ | $\epsilon_t$ | $1/(2\epsilon_t)$ | $1/(2(1-\epsilon_t))$ | $\alpha_t$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1/3 | 1/3 | 1/3 | +1 | +1 | +1 | 1/3 | 3/2 | 3/4 | 0.3466 |
| 2 | 1/4 | 2/4 | 1/4 | +1 | −1 | −1 | 1/4 | 2 | 2/3 | 0.5493 |
| 3 | 1/6 | 2/6 | 3/6 | −1 | −1 | +1 | 1/6 | 3 | 3/5 | 0.8047 |

**(b)** The final output scores for the instances are the following:

$$H(x_1) = +0.3466 + 0.5493 - 0.8047 = +0.0912$$
$$H(x_2) = +0.3466 - 0.5493 - 0.8047 = -1.0074$$
$$H(x_3) = +0.3466 - 0.5493 + 0.8047 = +0.6020$$

which after thresholding at 0 give the correct predictions $(+1, -1, +1)$.

**Exercise 2. Bagging of decision stumps**

Consider the bagging algorithm applied on the above data $D$, ensemble size $T = 3$, and the decision stump learner $\mathcal{A}$ as defined in Exercise 1.

**(a)** List all possible prediction vectors $(\hat{y}_1, \hat{y}_2, \hat{y}_3)$ that individual trees in the bagging ensemble might provide as predictions on the training data.

**(b)** *Note that this subtask involves combinatorics that we don't assume you know for Test 3.* What is the probability that the majority vote of the obtained ensemble gives correct predictions to all instances? Hint: for this first calculate the probability of each of the outcomes in subtask (a).

**Solutions of (a) and (b) combined.**

Bagging randomly samples 3 instances from $D$ with replacement. Keeping the order of picking there are $3 \cdot 3 \cdot 3 = 27$ possibilities. Among these is the one where $x_1$ gets picked each time, which can be represented as the weights $(3, 0, 0)$ for the instances $x_1, x_2, x_3$, respectively. Three out of 27 have $x_1$ picked twice and $x_2$ once, represented by weights $(2, 1, 0)$. Six out of 27 have $x_1, x_2, x_3$ each picked once, i.e. with weights $(1, 1, 1)$. As discussed in Exercise 1, the decision stump learner predicts all positives for such dataset with uniform weights. It also predicts all positives if the only negative instance $x_2$ does not get picked at all, this happens with weights $(3, 0, 0)$, $(2, 0, 1)$, $(1, 0, 2)$ and $(0, 0, 3)$. Overall, the probability of predicting all positives is $14/27$, because it happens for the 6 cases with weights $(1, 1, 1)$, 3 cases with $(2, 0, 1)$, 3 cases with $(1, 0, 2)$ and for the cases $(3, 0, 0)$ and $(0, 0, 3)$.

In the other extreme, if the bootstrapped dataset has no positives, then the model predicts all negatives. This happens with probability $1/27$, for the case $(0, 3, 0)$. In all remaining cases the weight for $x_2$ is non-zero, and at least one of the weights for $x_1$ and $x_3$ is zero. In such situation the instances with non-zero weights can be perfectly classified by the decision stump. For the 3 cases with weights $(2, 1, 0)$ and for the 3 cases with weights $(1, 2, 0)$ the prediction is $+--$. This happens with probability $6/27$. With the same probability the weights are either $(0, 2, 1)$ or $(0, 1, 2)$ and the prediction is $--+$.

We denote these 4 mutually exclusive events as $A, B, C, D$ and these can be summarised as follows:

| Event | Prediction | Probability | Cases |
|---|---|---|---|
| A | $+++$ | $14/27$ | $(1,1,1), (3,0,0), (2,0,1), (1,0,2), (0,0,3)$ |
| B | $---$ | $1/27$ | $(0,3,0)$ |
| C | $+--$ | $6/27$ | $(2,1,0), (1,2,0)$ |
| D | $--+$ | $6/27$ | $(0,1,2), (0,2,1)$ |

For the majority vote of $T = 3$ models to be correct on all instances we need at least 2 positive predictions out of 3 for instances $x_1$ and $x_3$ and at most 1 for $x_2$. Altogether this means at least 4 positive predictions across 3 models. Therefore, at least one of the models has to come from event A. As the number of positive predictions on instance $x_1$ has to be larger than on $x_2$, we need a model from event C. Similarly, we need a model from event D to have more positive votes on $x_3$ than on $x_2$.

From the above we can conclude that in order for the ensemble to be correct on all 3 instances we need models from events A,C,D. By fixing the order of models we need either ACD, ADC, CAD, CDA, DAC or DCA and each of those happens with probability $14/27 \cdot 6/27 \cdot 6/27$. Altogether, the probability is $6 \cdot 14 \cdot 6 \cdot 6/27^3 = 112/729 \approx 0.1536$.