

Ensemble methods (scored tasks: 4 points)

Tasks (Labs 8-9):

1. Implementation of AdaBoost

- (a) Implement your own version of AdaBoost algorithm. You can use available implementations of decision trees (or other base learners). Below is a description of AdaBoost algorithm:

Training:

- i. Define initial weights: $w_i = \frac{1}{n}$, $i = 1, \dots, n$.
- ii. For $k = 1, \dots, B$:
 - A. Build classifier f_k with weights: w_i .
 - B. Compute weighted classification error for k -th classifier:

$$\epsilon_k := \sum_{i=1}^n I[f_k(x_i) \neq y_i] w_i.$$

- C. Compute scaling factor: $\beta_k := \frac{\epsilon_k}{1-\epsilon_k}$.
- D. Add a pair (f_k, β_k) to the ensemble.
- E. For $i = 1, \dots, n$:
If $f_k(x_i) = y_i$ (example is correctly classified) then $w_i := w_i \beta_k$ (decrease weights).
- F. Normalize weights: $w_i = w_i / \sum(w_i)$.

Prediction:

- Prediction for new observation x :

$$\hat{y}(x) = \arg \max_y \sum_{k=1}^B I[f_k(x) = y] \log\left(\frac{1}{\beta_k}\right)$$

- Weighted voting.
- Classifier with small errors (small value of β_k) have larger weights in majority voting.

2. (a) Compare the following ensemble methods:

- Single tree
- Bagging
- Boosting (AdaBoost), your implementation.
- XGBoost
- Random Forest

- (b) Consider one real dataset (you can choose it from e.g. UCI repository) and one artificial dataset generated as follows:

- Generate $X_1, X_2, \dots, X_{10} \sim N(0, 1)$.
- Denote by $\chi_{10}^2(0.5)$ median of chi squared distribution with 10 degrees of freedom.
- Set $Y = 1$ if $\sum_{j=1}^{10} X_j^2 > \chi_{10}^2(0.5)$ and $Y = -1$ otherwise.
- Generate training data of size 2000 and testing data of size 10000.

- (c) Make a plot showing how the error changes with the number of iterations/trees (for boosting, bagging, Random Forest, XGBoost).