Algorithmic Data Analysis

Esther Galbrun

Spring 2024

UNIVERSITY OF
EASTERN FINLAND

Last updated on 2023-12-28 11:36

Q1.1: Multi-class running times

Suppose that a classification training algorithm requires $O(n^r)$ time for training on a data set of size n. Here r is assumed to be larger than 1. Consider a data set with a perfectly even distribution across k different classes.

As an example case, consider $r = 3$, $n = 1000$ and $k = 5$.

Compare the running time of the one-against-rest approach with that of the one-against-one approach.

Q1.2: Unbalance

Compare the characteristics of *resampling* vs. *reweighting* when dealing with unbalanced datasets.

efficiency

flexibility

out of the box

randomness

Q1.3: Boosting

Boosting is a $\left\{ \begin{array}{r} \text{data-centered} \\ \text{model-centered} \end{array} \right\}$ ensemble learning approach

aiming to reduce $\left\{ \begin{array}{r} \text{bias} \\ \text{variance} \end{array} \right\}$.

Q1.4: Bag and boost

Would you rather use a *linear SVM* or a *kernel SVM*?

i) To create the ensemble components in *bagging*

ii) To create the ensemble components in *boosting*