Group S: Alexander Wichmann, Artur Titkov, Rune Basedow

Task 1: Random Forests

(a)

Both, random forests and bagging, are ensemble methods which combine models to achieve more accurate results in their classifications, e.g. by avoiding problems such as overfitting. Bagging works by sampling (with replacement) from the original training set for each iteration. Then a classifier model is built in every iteration. The results of the different models are subsequently combined into one classification model (each one being weighted equally). The random forests method uses many decision tree classifiers in combination. This method basically starts with bagging to sample from the training set, but additionally uses a random selection of the attributes at each node to determine the split in every individual decision tree. Again, the final classification of unknown instances is then done by counting the result of every individual tree as one vote of equal weight and returning the class assigned by the majority of the decision trees.

(b) and (c)

Generally speaking, a higher number of trees leads to more accurate classification results. The strongest improvements of the results are made within the range of 1 to 10 trees used for the random forest. Whilst the 1 tree approach leads to an accuracy of around 83 percent, this rises to 89 percent and 91 percent for 5 and 10 trees respectively. Changing the depth of one individual decision tree (for exercise sheet 2) only lead to improved results up to a depth limited by the number of attributes that were available. Changing the number of decision trees used for the random forest, however, also improves results for considerably higher numbers of trees. This is due to the ensemble approach which avoids overfitting even better through a higher number of trees, even when only considering a limited number of attributes.

Task 3: Boosting

The use of strong learners limits the positive effects of boosting because the weighting is then only adjusted to a small degree.

Task 4: Bootstrap

The bootstrap method samples the training instances uniformly with replacement. When given a set with n instances, it is sampled n times with replacement. Therefore, the bootstrap sample will have n elements with some instances from the original data potentially occurring more than once. On average 63.3 percent of the original instances are sampled when using this approach, which leads to the name “0.632 bootstrap”.

Since every instance is chosen with probability 1/n and not chosen with probability 1-1/n and there are n selections, the probability of not including an instance is $(1 – 1/n)^n$. For large n this converges to 0.368 (which is e^{-1}), or 1 – 0.632.