Assignment 9: Ensemble – Theoretical Exercises for Blackboard Solutions

Machine Learning – University of Wrocław

Winter Semester 2025/2026

Note: The following exercises require understanding of theoretical material from lectures. Solutions should include mathematical derivations, proofs, or detailed calculations. These exercises are intended to be solved at the blackboard.

Task 1: Ensemble Success Probability (Bagging)

We have 101 independent binary classifiers, each with accuracy equal to 60%. We use majority voting to make the final decision.

(a) Derive the formula for the probability of correct classification by the ensemble.

(b) **Estimate** this probability for the given data.

(c) What is the minimum accuracy of a single classifier for the ensemble to be better than random guessing?

Task 2: Gini Impurity vs Entropy

A decision tree node contains 100 examples: 70 from class A and 30 from class B.

(a) Calculate the Gini Impurity for this node.

(b) Calculate the Entropy for this node.

(c) Consider a split into two child nodes:

- Left: 60 from class A, 10 from class B
- Right: 10 from class A, 20 from class B

Calculate the Information Gain for this split (using entropy).

(d) Explain when Gini and Entropy can give different splits.

Task 3: Classifier Weight in AdaBoost

In the AdaBoost algorithm, after training a weak learner, we calculate its weight as:

$$\alpha_m = \frac{1}{2} \ln \left(\frac{1 - \epsilon_m}{\epsilon_m} \right)$$

where ϵ_m is the weighted classification error.

(a) Prove that for $\epsilon_m = 0.5$, the weight $\alpha_m = 0$.

(b) Prove that for $\epsilon_m \to 0$, the weight $\alpha_m \to +\infty$.

(c) What happens when $\epsilon_m > 0.5$? Explain the intuition.

(d) A classifier has error $\epsilon = 0.3$. Calculate its weight α.

Task 4: Bootstrap and Out-of-Bag

In the Bagging (Bootstrap Aggregating) method, we sample n examples with replacement from the original set of n examples.

(a) What is the probability that a given example does NOT end up in the bootstrap sample?

(b) Calculate the limit of this probability as $n \to \infty$.

(c) What percentage of data on average constitutes the Out-of-Bag (OOB) set?

(d) Explain how OOB can serve as an estimator of generalization error.

Task 5: Gradient Boosting – Pseudo-residuals

In Gradient Boosting, each subsequent model learns to predict pseudo-residuals:

$$r_{im} = - \left[\frac{\partial L(y_i, F(\boldsymbol{x}_i))}{\partial F(\boldsymbol{x}_i)} \right]_{F=F_{m-1}}$$

(a) For the MSE loss function: $L(y, F) = \frac{1}{2}(y - F)^2$, show that pseudo-residuals are ordinary residuals.

(b) For the MAE loss function: $L(y, F) = |y - F|$, calculate the pseudo-residuals.

(c) Explain why MAE is more robust to outliers than MSE.

(d) For log-loss in binary classification: $L(y, F) = \log(1 + e^{-yF})$, where $y \in \{-1, +1\}$, derive the formula for pseudo-residuals.
