Ensemble Learning Theory

Outline

Introduction to Ensemble Learning

- **Ensemble Learning** combines multiple models to improve performance.
- Motivation:
  - Reduce overfitting
  - Improve generalization
  - Handle complex tasks
- Widely used in practice (e.g., Random Forest, XGBoost, Kaggle solutions)

Bias-Variance Decomposition

- Error = Bias^2 + Variance + Irreducible Error
- Ensembles reduce variance (Bagging), sometimes bias (Boosting)
- Averages out noise and instability

$$\text{Error} = \underbrace{\text{Bias}^2}_{\text{underfitting}} + \underbrace{\text{Variance}}_{\text{overfitting}} + \text{Noise}$$

Theoretical Justification

- Ensemble Error: depends on base learners' error and their correlation
- Key idea: uncorrelated models with decent accuracy help each other

$$E = \rho \cdot \bar{e} + (1 - \rho) \cdot \bar{e}^2$$

- $\bar{e}$: average individual error
- ρ: average correlation between classifiers

Diversity in Ensembles

- Diversity is critical for ensemble success
- Common diversity measures:
  - Q-statistic
  - Correlation coefficient
  - Disagreement measure
- Trade-off: Accuracy vs. Diversity

Bagging (Bootstrap Aggregating)

- Train each model on a bootstrap sample
- Combine predictions (voting/averaging)
- Reduces variance

$$\hat{f}_{bag}(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

- Example: Random Forests

Boosting

- Models are trained sequentially
- Each model focuses on errors of previous ones
- Reduces bias

$$F_m(x) = F_{m-1}(x) + \alpha_m h_m(x)$$

- AdaBoost: Changes sample weights
- Gradient Boosting: Fits gradient of loss function

Stacking

- Combine predictions using a meta-learner
- Level-0: Base learners
- Level-1: Meta model
- Trained on out-of-fold predictions

Conditions for Effective Ensembles

- Base learners should:
  - Be accurate (better than random)
  - Be diverse (make different errors)
- Independence isn't required — complementarity is enough

Advanced: PAC Learning Bounds

- PAC theory gives error bounds on learning
- Ensemble generalization error can be bounded:

$$R(H) \leq \hat{R}(H) + \sqrt{\frac{VC(H)\log(n)}{n}}$$

- Boosting improves margins $\rightarrow$ better generalization

Pruning and Bayesian View

- Ensemble pruning removes weak or redundant learners
- Bayesian model averaging:

$$p(y|x) = \sum_i p(y|x, h_i)p(h_i)$$

- Gives a probabilistic view of ensemble prediction

Practical Considerations

- Pros:
  - Higher accuracy
  - Robust to noise
- Cons:
  - Slower training/inference
  - Reduced interpretability
- Don't ensemble identical models!

Case Studies Tools

- Random Forest: Scikit-learn
- XGBoost, LightGBM: Efficient boosting libraries
- Deep Ensembles: Snapshot ensembles, SWA, MC Dropout
- Kaggle usage and industry adoption

Summary and Key Takeaways

- Ensembles improve performance by combining weak learners
- Success depends on diversity and accuracy
- Bagging, Boosting, and Stacking are key methods
- Balance performance and complexity

Thank you!