Ensemble methods

Gianluca Campanella

Ensemble methods

Idea

- Train a set of learners
- Combine their predictions

Motivation

- More flexible → reduce bias
- Less prone to overfitting → reduce variance

Why do ensembles work?

Assume 10 classifiers…

- Each with error rate ε = 30%
- Independent (not feasible in practice!)

What’s the probability that the ensemble makes a wrong prediction?

Why do ensembles work?

Assume 10 classifiers…

- Each with error rate ε = 30%
- Independent (not feasible in practice!)

What's the probability that the ensemble makes a wrong prediction?

- Wrong prediction $\rightarrow$ majority (≥ 6) makes a wrong prediction
- The probability is given by the binomial distribution,

$$\sum_{k=6}^{10} \binom{10}{k} \varepsilon^k (1-\varepsilon)^{10-k} \approx 4.73\% \ll \varepsilon$$

Contents

Bagging

Boosting

Stacking

Bagging

Bagging (bootstrap aggregating)

- Multiple datasets generated by sampling with replacement
- The probability that a given observation is *not* selected is

$$\left(1 - \frac{1}{n}\right)^n$$

- For large n, we have

$$\lim_{n \to +\infty} \left(1 - \frac{1}{n}\right)^n = \frac{1}{e} \approx 36.79\%$$

→ Each dataset contains slightly less than 2/3 of the observations

Bagging (bootstrap aggregating)

Variations

- Different types of learners
- Sampling of variables
- Sampling without replacement
- Size of subset

Boosting

Boosting

Idea

- ‘Later’ learners focus on observations that were predicted incorrectly by ‘earlier’ learners
- Aggregation is weighted by errors

Implementation

- Iterative procedure with weight updates
- Increase the weight of incorrectly classified observations

Boosting

Compared to bagging, boosting is…

- Much less noise-tolerant
- More accurate (if it works)
- Less well-calibrated

Stacking

Stacking

Idea

- Use a learner to combine predictions from multiple learners
- Trained using:

  $\mathbf{X}$ Cross-validated predictions

  $\vec{y}$ Class labels of training data

Meta-learners

- Not restricted to being linear (as in voting)
- Should learn how the (base) learners make mistakes