Bagging Classifiers

[Breiman, ML journal, 1996]
Bagging = Bootstrap aggregating
Bootstrap sampling: given set $D$ containing $m$ training examples
  - Create $D^i$ by drawing $m$ examples at random with replacement from $D$
  - $D^i$ expected to leave out $.37$ of examples from $D$
Bagging:
  - Create $k$ bootstrap samples $D^1, \ldots, D^k$
  - Train distinct classifier on each $D^i$
  - Classify new instance by classifier vote (equal weights)

Bagging - When Should this Help?

When learner is unstable
  - Learner is unstable if small change to training set causes large change in output hypothesis (e.g., decision trees, neural networks, not $k$ nearest neighbor)
  - Experimentally, bagging can help substantially for unstable learners, can somewhat degrade results for stable learners

Bagging Experiment

[Breiman, ML journal, 1996]
Given sample $S$ of labeled data, do 100 times and report average
1. Divide $S$ randomly into test set $T$ (10%) and training set $D$ (90%)
2. Learn decision tree from $D$.
   - $e_s \leftarrow$ error of tree on $T$
3. Do 50 times: Create bootstrap set $D^i$, learn decision tree, prune using $D$.
   - $e_B \leftarrow$ error of majority vote using trees to classify $T$
Bagging

Consider real valued target function, use mean instead of majority vote to combine classifier outputs

\[ \phi_A(x) = E_D \phi(x, D) \]

where \( \phi(x, D) \) is base classifier, \( \phi_A \) is aggregated classifier

\[ E_D(y - \phi(x, D))^2 = y^2 - 2yE_D \phi(x, D) + E_D \phi^2(x, D) \]

now using \( E_D \phi(x, D) = \phi_A(x) \), and \( EZ^2 \geq (EZ)^2 \),

\[ E_D(y - \phi(x, D))^2 \geq (y - \phi_A(x))^2 \]

so we expect a lower error for the bagged predictor \( \phi_A \).

---

Weighted Majority

[Relative mistake bound for Weighted-Majority] Let \( D \) be any sequence of training examples, let \( A \) be any set of \( n \) prediction algorithms, and let \( k \) be the minimum number of mistakes made by any algorithm in \( A \) for the training sequence \( D \). Then the number of mistakes over \( D \) made by the Weighted-Majority algorithm using \( \beta = \frac{1}{2} \) is at most

\[ 2.4(k + \log_2 n) \]