Ensemble Learning

- **Goal**
  - Improve accuracy of supervised learning task

- **Approach**
  - Use an ensemble of learners, rather than just one

- **Challenges**
  - How to construct ensemble
  - How to use individual hypotheses of ensemble to produce a classification
Ensembles of Classifiers

- Given ensemble of $L$ classifiers $h_1, \ldots, h_L$

- Decisions based on combination of individual $h_\ell$
  - E.g., weighted or unweighted voting

- How to construct ensemble whose accuracy is better than any individual classifier?
Ensembles of Classifiers

- Ensemble requirements
  - Individual classifiers disagree
  - Each classifier’s error < 0.5
  - Classifiers’ errors uncorrelated

- THEN, ensemble will outperform any $h_\ell$
Ensembles of Classifiers

P(ℓ of 21 hypotheses errant)  
Each hypothesis has error 0.3  
Errors independent  
P(11 or more errant) = 0.026
Boosting

- Each of $m$ training examples weighted according to classification difficulty $p_l(x)$
  - Initially uniform: $1/m$
- Training sample of size $m$ for iteration $l$ drawn with replacement according to distribution $p_l(x)$
- Learner biased toward higher-weight training examples – if learner can use $p_l(x)$
- Error $\varepsilon_l$ of classifier $h_l$ used to bias $p_{l+1}(x)$
- Learn $L$ classifiers
  - Each used to modify weights for next learned classifier
- Final classifier a weighted vote of individual classifiers
AdaBoost

Input: a set $S$, of $m$ labeled examples: $S = \{(x_i, y_i), i = 1, 2, \ldots, m\}$,
labels $y_i \in Y = \{1, \ldots, K\}$
LEARN (a learning algorithm)
a constant $L$.

[1] initialize for all $i$: $w_1(i) := 1/m$
[2] for $\ell = 1$ to $L$ do
[3] for all $i$: $p_\ell(i) := w_\ell(i) / (\sum_i w_\ell(i))$
[4] $h_\ell := \text{LEARN}(p_\ell)$
[5] $\epsilon_\ell := \sum_i p_\ell(i)[h_\ell(x_i) \neq y_i]$
[6] if $\epsilon_\ell > 1/2$ then
[7] $L := \ell - 1$
[8] goto 13
[9] $\beta_\ell := \epsilon_\ell / (1 - \epsilon_\ell)$
[10] for all $i$: $w_{\ell+1}(i) := w_\ell(i) \beta_\ell^{-1}[h_\ell(x_i) \neq y_i]$
[11] end for
[12] end for

[13] Output: $h_f(x) = \arg\max_{y \in Y} \sum_{\ell=1}^{L} \left( \log \frac{1}{\beta_\ell} \right)[h_\ell(x) = y]$
C4.5 with/without Boosting

Each point represents 1 of 27 test domains.
Summary

- **Advantages**
  - Ensemble of learners typically outperforms any one learner

- **Disadvantages**
  - Difficult to measure correlation between classifiers from different types of learners
  - Learning time and memory constraints
  - Learned concept difficult to understand