CPSC 340:
Machine Learning and Data Mining

Ensemble Methods

Bonus slides

Why can Voting Work?

- Consider 3 binary classifiers, each independently correct with probability 0.80:

- With voting, ensemble prediction is correct if we have "at least 2 right":
  - P(all 3 right) = 0.8^3 = 0.512.
  - P(2 rights, 1 wrong) = $3*0.8^2(1-0.8)$ = 0.384.
  - P(1 right, 2 wrongs) = $3*(1-0.8)^2 0.8$ = 0.096.
  - P(all 3 wrong) = $(1-0.8)^3$ = 0.008.
  - So ensemble is right with probability 0.896 (which is 0.512+0.384).

- Notes:
  - For voting to work, errors of classifiers need to be at least somewhat independent.
  - You also want the probability of being right to be > 0.5, otherwise it can do much worse.
  - Probabilities also shouldn't be too different (otherwise, it might be better to take most accurate).

Why does Bootstrapping select approximately 63%?

- Probability of an arbitrary x_i being selected in a bootstrap sample:

p(selected at least once in 'n' trials)

$= 1 -$ p(not selected in any of 'n' trials)

$= 1 - ($p(not selected in one trial)$)^n$

$= 1 - (1 - 1/n)^n$

$\approx 1 - 1/e$

≈ 0.63

(trials are independent)

(prob $= \frac{n-1}{n}$ for choosing any of the n-1 other samples)

($(1-1/n)^n \to e^{-1}$ as $n \to \infty$)

Why Averaging Works

- Consider ‘k’ independent classifiers, whose errors have a variance of σ^2.
- If the errors are IID, the variance of the vote is σ^2/k.
  - So the more classifiers that vote, the more you decrease error variance.
    (And the more the training error approximates the test error.)
- Generalization to case where classifiers are not independent is:

$$c\sigma^2 + \frac{(1-c)}{k}\sigma^2$$

  - Where ‘c’ is the correlation.
- So the less correlation you have the closer you get to independent case.
- Randomization in random forests decreases correlation between trees.
  - See also “Sensitivity of Independence Assumptions”.