Machine Learning IV
Ensembles

CSE 473

Machine Learning Outline

- Machine learning:
- Supervised learning
- Overfitting
- Ensembles of classifiers
  Bagging
  Cross-validated committees
  Boosting
  Stacking

Ensembles of Classifiers

- Assume
  Errors are independent
  Majority vote
- Probability that majority is wrong...
  = area under binomial distribution
  If individual area is 0.3
  Area under curve for ≥11 wrong is 0.026
  Order of magnitude improvement!

Constructing Ensembles

Bagging

- Generate k sets of training examples
- For each set
  Draw m examples randomly (with replacement)
  From the original set of m examples
- Each training set corresponds to
  63.2% of original
  (+ duplicates)
- Now train classifier on each set

Voting

Cross-validated committees

- Partition examples into k disjoint equiv classes
- Now create k training sets
  Each set is union of all equiv classes except one
  So each set has (k-1)/k of the original training data
- Now train a classifier on each set

Ensemble Construction II
**Ensemble Construction II**

**Cross-validated committees**

- Partition examples into $k$ disjoint equiv classes
- Now create $k$ training sets
  
  Each set is union of all equiv classes **except one**
  
  So each set has $(k-1)/k$ of the original training data

- Now train a classifier on each set

**Ensemble Creation III**

**Boosting**

- Maintain prob distribution over set of training ex
- Create $k$ sets of training data iteratively:
  - On iteration $i$
    
    Draw $m$ examples randomly (like bagging)
    
    But use probability distribution to bias selection
    
    Train classifier number $i$ on this training set
    
    Test partial ensemble (of $i$ classifiers) on all training exs
    
    Modify distribution: increase $P$ of each error ex

- Create harder and harder learning problems...

  "Bagging with optimized choice of examples"

**Ensemble Creation IV**

**Stacking**

- Train several base learners
- Next train meta-learner
  
  Learns when base learners are right / wrong
  
  Now meta learner arbitrates

  Train using cross validated committees
  
  - Meta-L inputs = base learner predictions
  
  - Training examples = 'test set' from cross validation