

Ensemble Learning

July 12, 2005
CS 486/686
University of Waterloo

Outline

- Ensemble Learning
 - Bagging
 - Boosting
- Reading: R&N Sect 18.4

CS486/686 Lecture Slides (c) 2005 P. Pruggert

2

Supervised Learning

- So far...
 - Decision trees
 - Statistical learning
 - Bayesian Learning
 - Maximum a posteriori
 - Maximum likelihood
 - Neural networks
- Which technique should we pick?

CS486/686 Lecture Slides (c) 2005 P. Pruggert

3

Ensemble Learning

- Sometimes each learning technique yields a different hypothesis
- But no perfect hypothesis...
- Could we combine several imperfect hypotheses into a better hypothesis?

CS486/686 Lecture Slides (c) 2005 P. Pruggert

4

Ensemble Learning

- Analogies:
 - Elections combine voters' choices to pick a good candidate
 - Committees combine experts' opinions to make better decisions
- Intuitions:
 - Individuals often make mistakes, but the "majority" is less likely to make mistakes.
 - Individuals often have partial knowledge, but a committee can pool expertise to make better decisions.

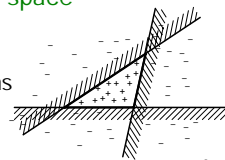
CS486/686 Lecture Slides (c) 2005 P. Pruggert

5

Ensemble Learning

- Definition: method to select and combine an ensemble of hypotheses into a (hopefully) better hypothesis

- Can enlarge hypothesis space
 - Perceptrons
 - linear separators
 - Ensemble of perceptrons
 - polytope

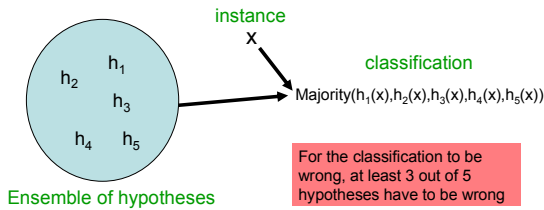


CS486/686 Lecture Slides (c) 2005 P. Pruggert

6

Bagging

- Majority Voting



CS486/686 Lecture Slides (c) 2005 P. Prasant

7

Bagging

- Assumptions:
 - Each h_i makes error with probability p
 - The hypotheses are independent
- Majority voting of n hypotheses:
 - k hypotheses make an error: $\binom{n}{k} p^k (1-p)^{n-k}$
 - Majority makes an error: $\sum_{k>n/2} \binom{n}{k} p^k (1-p)^{n-k}$
 - With $n=5, p=0.1 \rightarrow \text{err}(\text{majority}) < 0.01$

CS486/686 Lecture Slides (c) 2005 P. Prasant

8

Weighted Majority

- In practice
 - Hypotheses rarely independent
 - Some hypotheses have less errors than others
- Let's take a weighted majority
- Intuition:
 - Decrease weight of correlated hypotheses
 - Increase weight of good hypotheses

CS486/686 Lecture Slides (c) 2005 P. Prasant

9

Boosting

- Most popular ensemble technique
- Computes a weighted majority
- Can "boost" a "weak learner"
- Operates on a weighted training set

CS486/686 Lecture Slides (c) 2005 P. Prasant

10

Weighted Training Set

- Learning with a weighted training set
 - Supervised learning \rightarrow minimize train. error
 - Bias algorithm to learn correctly instances with high weights
- Idea: when an instance is misclassified by a hypotheses, increase its weight so that the next hypothesis is more likely to classify it correctly

CS486/686 Lecture Slides (c) 2005 P. Prasant

11

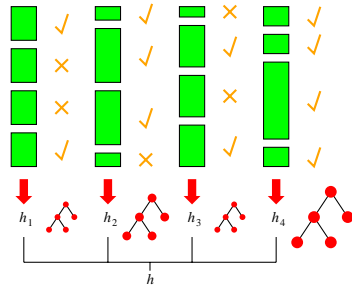
Boosting Framework

- Set all instance weights w_x to 1
- Repeat
 - $h_i \leftarrow \text{learn}(\text{dataset}, \text{weights})$
 - Increase w_x of misclassified instances x
- Until sufficient number of hypotheses
- Ensemble hypothesis is the weighted majority of h_i 's with weights w_i proportional to the accuracy of h_i

CS486/686 Lecture Slides (c) 2005 P. Prasant

12

Boosting Framework



CS486/686 Lecture Slides (c) 2005 P. Prasant

13

AdaBoost (Adaptive Boosting)

- $w_j \leftarrow 1/N \quad \forall_j$
- For $m=1$ to M do
 - $h_m \leftarrow \text{learn}(\text{dataset}, w)$
 - $\text{err} \leftarrow 0$
 - For each (x_j, y_j) in dataset do
 - If $h_m(x_j) \neq y_j$ then $\text{err} \leftarrow \text{err} + w_j$
 - For each (x_j, y_j) in dataset do
 - If $h_m(x_j) = y_j$ then $w_j \leftarrow w_j \text{err} / (1-\text{err})$
 - $w \leftarrow \text{normalize}(w)$
 - $z_m \leftarrow \log [(1-\text{err}) / \text{err}]$
- Return *weighted-majority*(h, z)

w: vector of N instance weights
z: vector of M hypoth. weights

CS486/686 Lecture Slides (c) 2005 P. Prasant

14

What can we boost?

- **Weak learner**: produces hypotheses at least as good as random classifier.
- Examples:
 - Rules of thumb
 - Decision stumps (decision trees of one node)
 - Perceptrons
 - Naive Bayes models

CS486/686 Lecture Slides (c) 2005 P. Prasant

15

Boosting Paradigm

- Advantages
 - No need to learn a perfect hypothesis
 - Can boost any weak learning algorithm
 - Boosting is very simple to program
 - Good generalization
- **Paradigm shift**
 - Don't try to learn a perfect hypothesis
 - Just learn simple rules of thumbs and boost them

CS486/686 Lecture Slides (c) 2005 P. Prasant

16

Boosting Paradigm

- When we already have a bunch of hypotheses, boosting provides a principled approach to combine them
- Useful for
 - Sensor fusion
 - Combining experts

CS486/686 Lecture Slides (c) 2005 P. Prasant

17

Boosting Applications

- Any supervised learning task
 - Spam filtering
 - Speech recognition/natural language processing
 - Data mining
 - Etc.

CS486/686 Lecture Slides (c) 2005 P. Prasant

18

Next Class

- Next Class:
 - Natural Language Processing
 - Russell and Norvig Ch. 22