Boosting algorithms

Boosting

– Main Assumption:
  - Combining many weak predictors to produce an ensemble predictor.

– Hypothesis Space
  - Variable size (nonparametric): Can model any function

Boosting (Continued)

- Each predictor is created by using a biased sample of the training data
  - Instances (training examples) with high error are weighted higher than those with lower error
- Difficult instances get more attention

Given: $(x_1, y_1), \ldots, (x_m, y_m)$ where $x_i \in X$, $y_i \in Y = \{-1, +1\}$
Initialize $D_1(i) = 1/m$.
For $t = 1, \ldots, T$:

Weights on training data

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \to \{-1, +1\}$ with error

$$\epsilon_t = \Pr_{i \sim D_t}\left[h_t(x_i) \neq y_i\right].$$

Training error

- Choose $\alpha_t = \frac{1}{2} \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$.
- Update:

This parameter weights the importance of the learning hypothesis h_t
Note that $\alpha_t \geq 0$ if $\epsilon_t \leq 1/2$ and that α_t gets larger as ϵ_t gets smaller.

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-\alpha_t} & \text{if } h_t(x_i) = y_i \\ e^{\alpha_t} & \text{if } h_t(x_i) \neq y_i \end{cases} = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

This rule increases the weight of examples misclassified by h_t

where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution).

Output the final hypothesis:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right).$$

The final hypothesis is a weighted majority vote of the T weak hypotheses

Figure 1: The boosting algorithm AdaBoost.

Learning:
Boosting and Adaboost

http://www.cs.ucsd.edu/~yfreund/adaboost/index.html

More slides here …boosting_and_bagging.pdf