Notes for Boosting

Shoufu Luo

November 14, 2013

1 Description

Weak learners, i.e. Naïve Bayes and Logistic Regression, have low variance but high bias, and therefore cannot solve hard learning problems. Can a set of weak classifiers create a single stronger learner?

AdaBoost, Adaptive Boosting, provides a framework to achieve this. Combining many weak classifiers that are good at different parts of the input space yield a stronger classifier. The idea is following: given a weak learner (slighter better than random guessing), run the learner by several iterations on re-weighted training data and generate one classifier for each iteration. In each iteration, each training instance will be re-weighted by how correctly it was classified by current iteration classifier, and used to train next-interation classifier. Then, let the learned classifiers vote and output the final classifier.

2 Algorithm

Given training data (x_i, y_i) where $i = 1, 2, ..., m$, and $x_i \in X, y_i \in Y = \{+1, -1\}$, let $D(i)$ the weight of *i-th* instance, and $D_t(i)$ the weight of *i-th* instance in *t-th* iteration. h_t[1] is the *t-th* iteration classifier.

Algorithm 1 Adaptive Boosting (AdaBoost)

Initialize $D_1(i) = \frac{1}{m}$ ▷ each instance has equal weight
for $t = 1 \to T$ **do**
    Find $h_t = \arg\min_{h_j \in H} \epsilon_j = \frac{1}{\sum_{i=1}^{m} D_t(i)} \sum_{i=1}^{m} D_t(i) \cdot \delta(y_i \neq h_j(x_i))$ ▷ indicator function $\delta(\cdot)$
    if $\epsilon_t \geq \frac{1}{2}$ **then** ▷ ϵ_j is the training error
        break
    end if
    $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$
    $Z_t = \sum_{i=1}^{m} D_t(i) \exp(-\alpha_t y_i h_t(x_i))$
    $D_{t+1}(i) = \frac{1}{Z_t} D_t(i) \exp(-\alpha_t y_i h_t(x_i))$ ▷ Update the weights for each instance
end for
$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$ ▷ Output the final Classifier

3 Facts

1. What is α_t and why $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$?

[1] All classifiers have the same form (base learner) but different parameters since they are trained by different weighted training data in each iteration

α_t is a strength for hypothesis h_t.

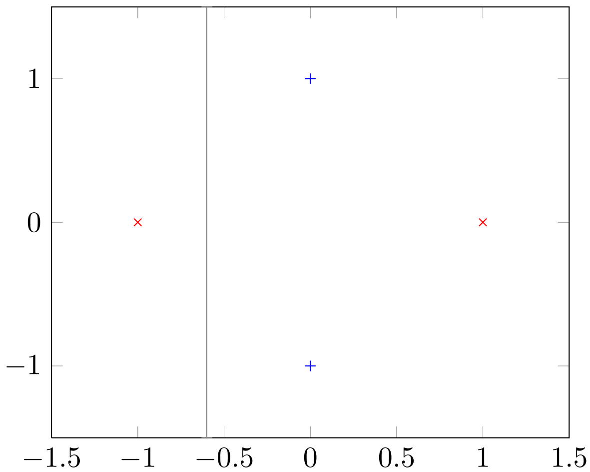
The training error of final classifier is bounded by: $\frac{1}{m}\sum_{i=1}^{m}\delta(H(x_i) \neq y_i) \leq \frac{1}{m}\sum_{i=1}^{m}\exp(-y_i f(x_i)) = \prod_t Z_t$, where $f(x) = \sum_t \alpha_t h_t(x)$ and $H(x) = sign(f(x))$.

If we minimize $\prod_t Z_t$, we minimize training error. Thus, we can tighten this bound greedily, by choosing α_t on each iteration to minimize each Z_t, which leads us to have $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$ [Freund & Schapire '97].

2. Prove $\frac{1}{m}\sum_{i=1}^{m}\exp(-y_i f(x_i)) = \prod_t Z_t$.
3. Boosting often is robust to overfitting (Not always). Test error decreases even after training error is zero.
4. Why $e^{-\alpha_t y_i h_t(x_i)}$?
5. How to learn h_t with ϵ_t?

4 Example

Consider the following toy dataset, consisting of 4 points, (0, -1, +), (1, 0, x), (-1, 0, x) and (0, 1, +), use decision stump as weak classifiers. Show how AdaBoost works, if set T=4. Each iteration, calculate ϵ_t, α_t, Z_t, and $D_t(i)$, and draw each weak classifier (e.g. h_1). Then, output the training error of AdaBoost.



5 Further Reading

[1] http://www.phillong.info/publications/LS10_potential.pdf
[2] http://en.wikipedia.org/wiki/Boosting_(meta-algorithm)
[3] http://en.wikipedia.org/wiki/AdaBoost
[4] http://www.cis.upenn.edu/~mkearns/teaching/COLT/adaboost.pdf