Boosted Classification

Yufei Chen

September 29, 2021

Supervised learning: mathematical formulation

- Let $X \in \mathbb{R}^p$ denote a real valued random input vector, and $Y \in \mathbb{R}$ a real valued random output variable, with joint distribution $\Pr(X, Y)$.
- Find the input-output relationship, that is, a function $f : \mathbb{R}^p \to \mathbb{R}$, such that

$$Y = f(X) + \epsilon$$

where ϵ captures measurements errors and other discrepancies.

Supervised learning: classification

For classification problems, the response variable Y is categorical (qualitative). And our goals are to:

- Build a classifier $\widehat{C}(X)$ that assigns a class label $\mathcal{C}$ to a future unlabeled observation X.
- Assess the uncertainty in each classification.
- Understand the roles of the different features among $X = (X_1, X_2, \ldots, X_p)$.

what is boosting?

- Boosting is a general ensemble method that can be applied to many statistical learning problems including classification.
- Boosting is a sequential ensemble method. In each iteration, boosting can involve resampling method to create multiple copies of the original training data.
- Boosting fits the learner targeting on the residuals or error produced by the previous iteration. And it combines all the learners to form the final one with satisfactory performance.

AdaBoost.M1

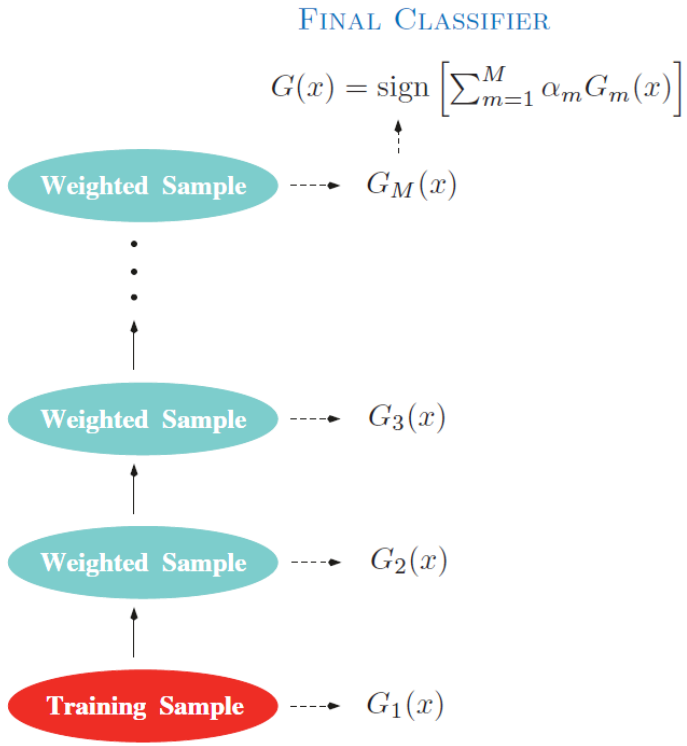
- Consider a two-class problem, with the output variable coded as $Y \in \{-1, 1\}$. Given a vector of predictor variables X, a classifier $C(X)$ produces a prediction taking one of the two values $\{-1, 1\}$. The error rate on the training sample is

$$\overline{\text{err}} = \frac{1}{N} \sum_{i=1}^{N} (y_i \neq C(x_i)),$$

  and the expected error rate on future predictions is $\text{E}_{XY} I(Y \neq C(X))$.
- The purpose of boosting is to sequentially apply the weak classification algorithm to repeatedly modified versions of the data, thereby producing a sequence of weak classifiers $C_m(x)$, $m = 1, 2, \ldots, M$.

Schematic of AdaBoost

- $\alpha_1, \alpha_2, \ldots, \alpha_M$ are computed by the boosting algorithm, and weight the contribution of each respective $G_m(x)$.
- The effect is to give higher influence to the more accurate classifiers in the sequence.



Algorithm 1: Adaboost.M1

1. Initialize the observation weights $w_i = 1/N$, $i = 1, \ldots, N$;
2. **for** $m = 1$ **to** M **do**
   (a) Fit a classifier $C_m(x)$ to the training data using weights w_i;
   (b) Compute
   $$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq C_m(x_i))}{\sum_{i=1}^{N} w_i}.$$
   (c) Compute $\alpha_m = \log\left((1 - err_m)/err_m\right).$;
   (d) Set $w_i \leftarrow w_i \cdot \exp\left[\alpha_m \cdot I(y_i \neq C_m(x_i))\right], i = 1, \ldots, N.$;

end

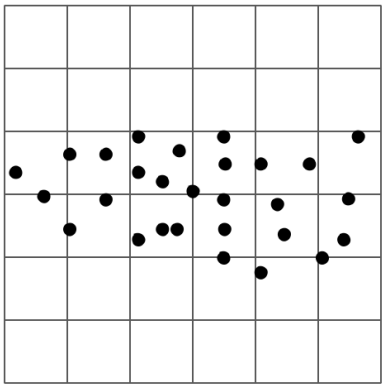
3. Output $C(x) = \text{sign}\left[\sum_{m=1}^{M} \alpha_m C_m(x)\right]$

Discussion about adaboost

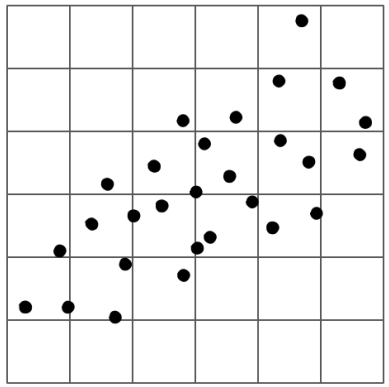
- The fundamental algorithm in boosted classification area. The majority of boosted classification algorithms are proposed based on the idea of the weights of samples and classifiers.
- Theoretical analysis on boosted classification are restricted in adaboost at present, new algorithms with the state-of-the-art performances are not theoretically armed.
- A theoretical analysis framework to boosted classification is in demand.

Our idea

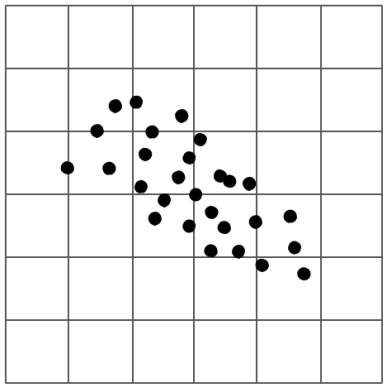
Histogram transform with partition



Original



Transformation induced by H1



Transformation induced by H2

In each iteration, we generate a histogram transform H_t including stretching, rotating and transforming.

Randomness

- Unlike traditional adaboost, we introduce randomness into the algorithm in each iteration, not only in the histogram transform, but also in the bootstrap resampling.
- We assume that the samples in the same cell of histogram partition have the same category. Literally, this rough estimation can be fit quickly and its performance can be greatly enhanced by boosting.
- Another reason to establish such algorithm derives from the theoretical properties.

Theoretical structures

- **Regularized boosting classifiers**:
  Consider a family $(h_i)_{i\in I} w_i^* h_i$ of weak classifiers $h_i : X \to \mathbb{R}$, and the weighted combination $f_{w^*} := \sum_{i\in I} w_i^* h_i$ is constructed.
- **An oracle inequality for regularized boosting algorithms**:

$$P^n\left(D : \lambda||f_{D,\lambda}||_E + \mathcal{R}_{L,P}(f_{D,\lambda}) - \mathcal{R}^*_{L,P} \geq 2A(\lambda) + \lambda\right) \leq e^{-\tau},$$

- **Another oracle inequality**:

$$\begin{aligned}&\lambda||\widehat{f}_{D,\lambda}||_E + \mathcal{R}_{L_\delta,P}(f_{D,\lambda}) - \mathcal{R}^*_{L_\delta,P}\\ &\quad< 15A(\lambda) + K\left(\frac{a^{2p}}{\lambda^{2p}n}\right)^{\frac{1}{1-p}} + K\frac{\tau}{n} + \frac{30A(\lambda)}{\lambda n} + 3\epsilon\end{aligned}$$