Error Correcting Output Codes for Multiclass Machine Learning

M. Krechetov

Department of Computer Science
Higher School of Economics

IITP RAS

Optimization and Statistical Learning Group weekly seminar,
2016
Outline

1. Introduction
   - Basics of coding theory
   - Multiclass problem
   - Reducing multiclass to binary

2. Boosting algorithms
   - AdaBoost
   - AdaBoost.ECC

3. Boosting as a gradient descent

4. Generalization bounds
Outline

1 Introduction
   ■ Basics of coding theory
     ■ Multiclass problem
     ■ Reducing multiclass to binary

2 Boosting algorithms
   ■ AdaBoost
   ■ AdaBoost.ECC

3 Boosting as a gradient descent

4 Generalization bounds
Definitions

- Channel is a string of 0s and 1s.
- Errors are independent Bernoulli variables $e_i$.
- Block code $f : \{0, 1\}^N \rightarrow \{0, 1\}^K$.
- Repetition code.
- Hamming decoding and distance.
- Error-correcting codes.
- Capacity of a channel $= \frac{N}{K}$. 
Shannon theorem

- Entropy function $H(p) = -p \cdot \log_2(p) - (1 - p) \cdot \log_2(1 - p)$.

Theorem (Shannon, 1948)

For any $\varepsilon > 0$ there is a block code with channel capacity greater than $1 - H(p) - \varepsilon$ and with overall error smaller than $\varepsilon$. 
Introduction

Multiclass problem

Reducing multiclass to binary

Boosting algorithms

AdaBoost

AdaBoost.ECC

Boosting as a gradient descent

Generalization bounds
Multiclass setting

- Sample: \( (x_i, y_i)_{i=1}^m \) with \( x_i \in X = \mathbb{R}^d \) and \( y_i \in Y = \{1, \ldots, k\} \).
- \( (x_i, y_i)_{1}^{m} \) are sampled according to a some distribution \( D \).
- Hypothesis is a function \( h : X \rightarrow Y \). Our goal is to minimize expected error.
Outline

1. Introduction
   - Basics of coding theory
   - Multiclass problem
   - Reducing multiclass to binary

2. Boosting algorithms
   - AdaBoost
   - AdaBoost.ECC

3. Boosting as a gradient descent

4. Generalization bounds
Reduction to binary

- One-vs-all scheme.
- Reduction with a code matrix of an error-correcting code \((n, k, d)\).
Some facts about training error and correlation

**Fact №1**

The worst-case training error of this scheme can be no higher than \( \frac{2n}{d} \) times the average error \( \frac{1}{n} \sum \epsilon_i \).

**Fact №2**

If \( \Delta \) is an upper bound on \( \mathbb{P}(h_i(x) \text{ is wrong and } h_j(x) \text{ is wrong}) \), then the worst-case training error can be no higher than \( 4 \frac{n(n-1)}{d(d-2)} \Delta \).
Outline

1. Introduction
   - Basics of coding theory
   - Multiclass problem
   - Reducing multiclass to binary

2. Boosting algorithms
   - AdaBoost
   - AdaBoost.ECC

3. Boosting as a gradient descent

4. Generalization bounds
Given $(x_i, y_i)_{i=1}^{m}$ with $x_i \in X$ and $y_i \in Y = \{-1, 1\}$.

Initialize $D_1(i) = \frac{1}{m}$.

For $t = 1, ..., T$:

- Train weak learner using distribution $D_t$.
- Get weak hypothesis $h_t : X \rightarrow \mathbb{R}$.
- Choose $\alpha_t$.

- Update:
  
  $$D_{t+1}(i) = \frac{D_t(i) \cdot \exp(-\alpha_t y_i h_t(x_i))}{Z_t},$$

  where $Z_t$ is a normalization factor.

Output the final hypothesis $H(x) = \text{sgn}(\sum_{t=1}^{T} \alpha_t h_t(x))$. 

---

**AdaBoost**
Outline

1. Introduction
   - Basics of coding theory
   - Multiclass problem
   - Reducing multiclass to binary

2. Boosting algorithms
   - AdaBoost
   - AdaBoost.ECC

3. Boosting as a gradient descent

4. Generalization bounds
Given \((x_i, y_i)^m_{i=1}\) with \(x_i \in X\) and \(y_i \in Y = \{1, \ldots, k\}\).

Initialize \(\overline{D}_1(i, l) = \frac{1}{m(k-1)}\) if \(l \neq y_i\) and \(\overline{D}_1(i, l) = 0\) otherwise.

For \(t = 1, \ldots, T:\)

- Compute coloring \(\mu_t : Y \rightarrow \{-1, 1\}\).
- Let \(U_t = \sum_{i=1}^{m} \sum_{l \in Y} \overline{D}_1(i, l)[\mu_t(l) \neq \mu_t(y_i)]\).
- Let \(D_t(i) = \frac{1}{U_t} \sum_{l \in Y} \overline{D}_1(i, l)[\mu_t(l) \neq \mu_t(y_i)]\).
- Get weak hypothesis \(h_t : X \rightarrow \{-1, 1\}\).
- Choose \(g_t(x) = \alpha_t\), if \(h_t(x) = 1\) and \(g_t = -\beta_t\) otherwise.
- Update:
  \[
  \overline{D}_{t+1}(i) = \frac{\overline{D}_t(i) \cdot \exp((g_t(x_i)\mu_t(l) - g_t(x_i)\mu_t(y_i))/2)}{Z_t},
  \]
  where \(Z_t\) is a normalization factor.

Output the final hypothesis \(H(x) = \arg \max_{l \in Y} \left(\sum_{t=1}^{T} g_t(x)\mu_t(l)\right)\).
Last time we have proved that in symmetric case ($\alpha_t = \beta_t$):

$$\hat{\epsilon} \leq (k - 1) \cdot \exp\left(\sum_{t=1}^{T} -2\gamma_t^2 U_t^2\right), \text{ where } \gamma_t = \frac{1}{2} - \epsilon_t$$
For Further Reading I

D. MacKay.

T. Dietterich., G. Bakiri

V. Guruswami., A. Sahai
*Multiclass Learning, Boosting, and Error-Correcting Codes*, 1999.

E. Allwein