Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning

Blanca Cano Camarero

Universidad Autónoma de Madrid

May 16, 2023

Overview

1. Article information
2. Generalization
3. Theoretical results

Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning

- Ali Rahimi and Recht
- Year 2008
- Cited by 720.

Rahimi and Recht (2008)

Objetive The main technical contributions of the paper are an approximation error bound (Lemma1), and a synthesis of known techniques from learning theory to analyze random shallow networks.

Learning for me and some observation

- Generalization of the bound for several shallow methods.
- Randomness in shallow architecture: What about depth ones?
- The results are applicable only for binary classification problem. Could be properly generalized or this kind of method are more suitable for binary classification of only for classification.

Introduction

These randomized shallow networks have largely been superceded by optimally, or nearly optimally, tuned shallow architectures such as weighted sums of positive definite kernels (as in Support Vector Machines), or weigted sums of weak classifiers (as in Adaboost). But recently, architectures that randomly transform their inputs have been resurfacing in the machine learning community

Classification problem

Consider the problem of fitting a function $f : X \longleftarrow \mathbb{R}$ to train a data set of m input-output pairs drawn iid from some unknown distribution $P(x, y)$ with $x \in X$ and $y \in \{-1, 1\}$.
That minimizes the empirical risk:

$$R_{emp[f]} = \frac{1}{m} \sum_{i}^{m} c\left(f(x_i), y_i\right). \tag{1}$$

Where c penalizes the deviation between the prediction f.

- Hinge loss for SVM,
- Exponential los in Adabosst,
- quadratic loss for Least squarest classification.

Form of f

$$f(x) = \sum_{i=1}^{\infty} \alpha(w_i)\phi(x; w_i), \tag{2}$$

or

$$f(x) = \int \alpha(w_i)\phi(x; w_i)dw \tag{3}$$

here feature functions $\phi : X\Omega \longrightarrow \mathbb{R}$ parameterized by some vector $w \in \Omega$, are weighted by a function $\alpha : \Omega \longrightarrow \mathbb{R}$.

- Kernel machines ϕ are eigenfunctions of a positive kernel k,
- Adaboost: decision trees.

Objective: minimization

$$\text{minimize}_{w_1,\ldots,w_k\in\Omega} \quad {}_{\alpha}R_{emp}\left[\sum_{k}^{K}\phi(x;w_k)\alpha_k.\right] \tag{4}$$

Idea: **Randomize over** w **and minimize over** α Crucial selection: p distribution from which w are drawn.

Algorithm

Algorithm 1: The Weighted Sum of Random Kitchen Sinks fitting procedure
Input: A dataset $\{x_i, y_i\}_{i=1}^{m}$ of m points, a bounded feature function $|\phi(x; w)| \leq 1$, an integer K, a scalar C, and a probability distribution $p(w)$ on the parameters of ϕ.
Output: A function $\hat{f}(x) = \sum_{k=1}^{K} \phi(x; w_k)\alpha_k$.
1. Draw $w_1, \ldots, w_K$ independently and identically distributed from p. 2. Featurize the input: $z_i \leftarrow [\phi(x_i; w_1), \ldots, \phi(x_i; w_K)]^T$. 3. With w fixed, solve the empirical risk minimization problem:

$$\text{minimize} \quad \alpha \in \mathbb{R}^K$$

$$\frac{1}{m} \sum_{i=1}^{m} c(\alpha^T z_i, y_i)$$

$$\text{subject to} \quad \|\alpha\|_\infty \leq \frac{C}{K}$$

Theoretical results

Formally, we show that the Algorithm 1 returns a function that has low true risk. The true risk of a function f is defined as:

$$R[f] \equiv \mathbb{E}_{(x,y)\sim P} c(f(x), y), \tag{5}$$

and measures the expected loss of f on as-yet-unseen test points, assuming these test points are generated from the same distribution that generated the training data. The following theorem states that with very high probability, Algorithm 1 returns a function whose true risk is near the lowest true risk attainable by functions in the class F_p defined below:

Main theorem

Theorem 1 (Main Result). Let p be a distribution on Ω, and let ϕ satisfy

$$\sup_{x,w} |\phi(x; w)| \leq 1. \tag{6}$$

Define the set

$$F_p \equiv \left\{ f(x) = \int_\Omega \alpha(w)\phi(x; w)\, dw \,\middle|\, |\alpha(w)| \leq Cp(w) \right\}. \tag{7}$$

Suppose $c(y, y') = c(yy')$, with $c(yy')$ being L-Lipschitz. Then, for any $\delta > 0$, if the training data $\{x_i, y_i\}_{i=1}^m$ are drawn independently and identically distributed (iid) from some distribution P, Algorithm 1 returns a function $\hat{f}$ that satisfies

$$R[\hat{f}] - \min_{f \in F_p} R[f] \leq \mathcal{O}\left(\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{K}} \right) LC \sqrt{\log \frac{1}{\delta}} \right), \tag{8}$$

with probability at least $1 - 2\delta$ over the training dataset and the choice of parameters $w_1, \ldots, w_K$.

References I

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL `https://proceedings.neurips.cc/paper_files/paper/2008/file/0efe32849d230d7f53049ddc4a4b0c60-Paper.pdf`.