IFT 6085 - Lecture 15
Weighted Sums of Random Kitchen Sinks: Replacing minimization with randomization in learning

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes **Instructor:** Ioannis Mitliagkas
Winter 2019: [Jonathan Guymont, Marzieh Mehdizadeh]

1 Summary

Consider the one hidden layer multilayer perceptron with identity output activation function $f(\mathbf{x}) = \mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x})$ where σ could be a non linear activation function. A standard way to ensure that f is a good mapping from the input $\mathbf{x} \in \mathcal{X}$ to the output $y \in \mathcal{Y}$ is to optimize (e.g. via SGD) $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ such that they minimize the empirical risk. Now consider drawing $\mathbf{W}^{(1)}$ from some distribution $p(\mathbf{W})$ and optimizing the empirical risk over $\mathbf{W}^{(2)}$ only. In this setup, we have $f(\mathbf{x}) = \mathbf{W}^{(2)}\phi(\mathbf{x}; \mathbf{W}^{(1)})$ where ϕ is a deterministic feature map that is initialized randomly. The authors in [1] showed that even if the parameter of the feature map are not optimized, minimizing the empirical risk with respect to $\mathbf{W}^{(2)}$ returns a function whose true risk is near the lowest true risk attainable by an infinite-dimensional class of functions $\mathcal{F}_p$ defined as below:

$$\mathcal{F}_p \equiv \left\{ f(x) = \int_\Omega \alpha(\omega)\phi(x;\omega)d\omega \;\Big|\; |\alpha(\omega)| \leq Cp(\omega) \right\} \tag{1}$$

where $p(\omega)$ is the distribution from which $\mathbf{W}^{(1)}$ was drawn.

2 Introduction

Given a set of training data in a domain $\{x^{(i)}, y^{(i)}\}_{i=1,\dots,m}$, $x^{(i)} \in \mathcal{X}$, $y^{(i)} \in \{-1, 1\}$ the goal is to learn the mapping $f \colon \mathcal{X} \mapsto \mathcal{Y}$ that minimizes the empirical risk

$$\hat{R}_S[f] = \sum_{(x,y)\in S} l\left(h(x), y\right) \tag{2}$$

where l is a loss function that specifying the penalty assign to the deviation between the prediction $f(x)$ and the ground truth y and $S \subset (\mathcal{X} \times \mathcal{Y})$.

Similarly to kernel machines, we will consider functions of the form

$$f(x) = \sum_i \alpha(\omega_i)\phi(x;\omega_i)\mathrm{d}\omega \tag{3}$$

if $\{\omega_i\}$ is a discrete set, or

$$f(x) = \int \alpha(\omega)\phi(x;\omega)\mathrm{d}\omega \tag{4}$$

if ω is continuous. The function $\phi\colon \mathcal{X} \mapsto \mathbb{R}$ is a feature map parametrized by some vector $\omega \in \Omega$ that are weighted by a function $\alpha\colon \Omega \mapsto \mathbb{R}$. Let $\boldsymbol{\omega}^*, \boldsymbol{\alpha}^*$ be the vectors of weights that minimize the empirical risk, i.e.

$$\boldsymbol{\omega}^*, \boldsymbol{\alpha}^* = \underset{\omega_1,\ldots,\omega_K \in \Omega,\ \alpha_1,..,\alpha_K \in \mathcal{A}}{\arg\min} \hat{\mathbf{R}}_S \left[\sum_{k=1}^{K} \phi(x;\omega_k)\alpha_k \right] \tag{5}$$

A standard approach in machine learning is to use some optimization procedure to approximate $\boldsymbol{\omega}^*$ and $\boldsymbol{\alpha}^*$. However, the authors propose less orthodox way approximate the empirical risk minimizer; instead of optimizing w.r.t $\boldsymbol{\omega}$ and $\boldsymbol{\alpha}$, draw $\boldsymbol{\omega}$ from some distribution $p(\boldsymbol{\omega})$ and optimize over $\boldsymbol{\alpha}$ only. Algorithm (1) describe the procedure.

Algorithm 1 Pseudocode for Anomaly detection

Input: A dataset $\{x^{(i)}, y^{(i)}\}_{i=1,\ldots,n}$
Input: A bounded feature function $|\phi(x;\omega)| \leq 1$
Input: $K \in \mathbb{N}$
Input: $C \in \mathbb{R}$
Input: A probability distribution $p(\boldsymbol{\omega})$
Output: A function $\hat{h}(x) = \sum_{k=1}^{K} \phi(x;\omega_k)\alpha_k$
Draw $\boldsymbol{\omega} \in \mathbb{R}^K$ from $p(\boldsymbol{\omega})$
Featurize the input: $\mathbf{z}^{(i)} \leftarrow \phi(\mathbf{x}^{(i)};\boldsymbol{\omega})$
With $\boldsymbol{\omega}$ fixed, solve the empirical risk minimization problem

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha} \in \mathbb{R}^K}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} l\left(\boldsymbol{\alpha}^\top \mathbf{z}^{(i)}, y^{(i)}\right) \tag{6}$$

s.t $||\boldsymbol{\alpha}||_\infty \leq C/K.$

The following theorem (1) states that algorithm (1) has low *true risk*. The true risk $\mathbf{R}[h]$ is defined as the expected loss on points drawn from the data distribution $\mathcal{D}$.

$$\mathbf{R}[f] = \mathbb{E}_{(x,y)\sim\mathcal{D}} l(f(x), y) \tag{7}$$

More specifically, theorem (1) states Algorithm (1) returns a function whose true risk is near the lowest true risk attainable by an infinite-dimensional class of functions $\mathcal{F}_p$ defined below:

Theorem 1. *(Main result) Let p be a distribution on Ω, and let ϕ satisfy $\sup_{x,w} |\phi(x;w)| \leq 1$ (uniformly bounded). Define the hypothesis set as follows:*

$$\mathcal{F}_p \equiv \left\{ f(x) = \int_\Omega \alpha(\omega)\phi(x;\omega)d\omega \;\middle|\; |\alpha(\omega)| \leq Cp(\omega) \right\} \tag{8}$$

Suppose the loss function is as below $l(y, y^{'}) = l(yy^{'})$, with $l(yy^{'})$ L-Lipschitz. Then for any $\delta > 0$, if the training data $\{x_i, y_i\}_{i=1\cdots m}$ are drawn $i.i.d$ from some distribution P, Algorithm 1 returns a function $\hat{f}$ that satisfies

$$\mathbf{R}[\hat{f}] - \min_{f \in \mathcal{F}_p} \mathbf{R}[f] \leq O\left\{ \left(\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{K}} \right) LC \log \sqrt{\log 1/\delta} \right) \right\}$$

with probability at least $1 - 2\delta$ over the training dataset and the choice of the parameters $\omega_1, \cdots, \omega_K$.

C is arbitrarily chosen and can be considered as a regulirizer. The hypothesis set $\mathcal{F}_p$ is quite rich. It consists of functions whose weights $\alpha(\omega)$ decays more rapidly than the given sampling distribution p.

3 Steps to prove the Main Theorem

Algorithm 1 returns a function that lies in the random set:

$$\hat{\mathcal{F}}_\omega \equiv \left\{ f(x) = \int_\Omega \alpha(\omega)\phi(x;\omega)d\omega \;\middle|\; |\alpha(\omega)| \leq C/K \right\}$$

We are going to see how much we loose by going from $\mathcal{F}_p$ to $\hat{\mathcal{F}}_\omega$.
The upper bound in the main theorem can be decomposed in a standard way into two bounds:

- An approximation error bound that shows that the lowest true risk attainable by a function in $\hat{\mathcal{F}}_\omega$ is not much larger than the lowest true risk attainable in $\mathcal{F}_p$ (Lemma 2).
- An estimation error bound that shows that the true risk of every function in $\hat{\mathcal{F}}_\omega$ is close to its empirical risk (Lemma 3)

The following Lemma is helpful in bounding the approximation error:

Lemma 1. *Let μ be a measure on $\mathcal{X}$, and f^* a function in $\mathcal{F}_p$. If $\omega_1, \cdots, \omega_K$ are drawn $i.i.d$ from p, then for any $\delta > 0$, with probability at least $1-\delta$ over $\omega_1, \cdots, \omega_K$, there exists a function $\hat{f} \in \mathcal{F}_\omega$ so that*

$$\sqrt{\int_\mathcal{X} \left(\hat{f}(x) - f^*(x)\right)^2 d\mu(x)} \leq \frac{C}{\sqrt{K}}\left(1 + \sqrt{2\log 1/\delta}\right)$$

Lemma 2. *(Bound on the approximation error) Suppose $l(y, y^{'})$ is L-Lipschitz in its first argument. Let f^* be a fixed function in $\mathcal{F}_p$. If $\omega_1, \cdots, \omega_K$ are drawn $i.i.d$ from p, then for any $\delta > 0$, with probability at least $1-\delta$ over $\omega_1, \cdots, \omega_K$, there exists a function $\hat{f} \in \hat{\mathcal{F}}_\omega$ that satisfies*

$$\mathbf{R}[\hat{f}] \leq \mathbf{R}[f^*] + \frac{LC}{\sqrt{K}}\left(1 + \sqrt{2\log 1/\delta}\right)$$

A standard result from statistical learning theory states that for a given choice of $\omega_1, \cdots, \omega_K$the empirical risk of every function in $\hat{\mathcal{F}}_\omega$ is close to its true risk. The following lemma can be proven by using Holder inequality.

Lemma 3. *(Bound on the estimation error). Suppose $l(y, y^{'}) = l(yy^{'})$, with $l(yy^{'})$ L-Lipschitz. Let $\omega_1, \cdots, \omega_K$ be fixed. If $\{x_i, y_i\} \quad i = 1 \cdots m$ are drawn $i.i.d$ from a fixed distribution, for any $\delta > 0$, with probability at least $1-\delta$ over the dataset, we have*

$$\forall \hat{f} \in \hat{\mathcal{F}}_\omega \quad \left|\mathbf{R}[f] - \hat{\mathbf{R}}[f]\right| \leq \frac{1}{\sqrt{m}}\left(4LC + 2|c(0)| + LC\sqrt{\frac{1}{2}\log 1/2}\right)$$

No we are ready to give a sketch of the proof of main theorem by using the above lemmas.

Proof of theorem 1. Let f^* be a minimizer of the true risk $\mathbf{R}$ over $\mathcal{F}_p$, $\hat{f}$ be a minimizer of the empirical risk $\hat{\mathbf{R}}$ over $\hat{\mathcal{F}}_\omega$ (i.e. $\hat{f}$ is the output of Algorithm 1), and $\hat{f}^*$ be a minimizer of the true risk $\mathbf{R}$ over $\hat{\mathcal{F}}_\omega$ (i.e. $\hat{f}^*$ is the optimal output of Algorithm 1). Then

$$\begin{aligned}\mathbf{R}[\hat{f}] - \mathbf{R}[f^*] =& \mathbf{R}[\hat{f}] - \mathbf{R}[\hat{f}^*] + \mathbf{R}[\hat{f}^*] - \mathbf{R}[f^*] && (9)\\ \leq& |\mathbf{R}[\hat{f}] - \mathbf{R}[\hat{f}^*]| + \mathbf{R}[\hat{f}^*] - \mathbf{R}[f^*] && (10)\end{aligned}$$

Let ϵ_{est} denote the upper bound of the right side of the inequality in Lemma 3:

$$\epsilon_{\text{est}} = \frac{1}{\sqrt{m}}\left(4LC + 2|c(0)| + LC\sqrt{\frac{1}{2}\log 1/2}\right)$$

With probability at least $1-\delta$ we have

$$\begin{aligned}
|\mathbf{R}[\hat{f}] - \mathbf{R}[\hat{f}^*]| =& |\mathbf{R}[\hat{f}] + \hat{\mathbf{R}}[\hat{f}^*] - \hat{\mathbf{R}}[\hat{f}^*] - \mathbf{R}[\hat{f}^*]| \\
\leq& |\mathbf{R}[\hat{f}] + \underbrace{\hat{\mathbf{R}}[\hat{f}^*] - \hat{\mathbf{R}}[\hat{f}]}_{\geq 0} - \mathbf{R}[\hat{f}^*]| \qquad \text{(By optimality of } \hat{f}\text{)} \\
\leq& |\mathbf{R}[\hat{f}] - \hat{\mathbf{R}}[\hat{f}]| + |\mathbf{R}[\hat{f}^*] - \hat{\mathbf{R}}[\hat{f}^*]| \\
\leq& 2\epsilon_{\text{est}} \qquad \text{(By Lemma 3)}
\end{aligned}$$

Let ϵ_{app} denote the right term in the upper bound of the inequality in Lemma 2:

$$\epsilon_{\text{app}} = \frac{LC}{\sqrt{K}}\left(1 + \sqrt{2\log 1/\delta}\right)$$

Also note that $\mathbf{R}[\hat{f}^*] < \mathbf{R}[\hat{f}]$ since $\hat{f}^*$ minimize the true risk over $\mathcal{F}_\omega$. Using this fact we have that with probability at least $1-\delta$ the following inequality hold

$$\begin{aligned}
\mathbf{R}[\hat{f}^*] - \mathbf{R}[f^*] \leq& \mathbf{R}[\hat{f}] - \mathbf{R}[f^*] \qquad (\hat{f}^* \text{ minimize } \mathbf{R} \text{ over } \mathcal{F}_\omega) \\
\leq& \epsilon_{\text{app}} \qquad \text{(Lemma 2)}
\end{aligned}$$

Hence

$$\mathbf{R}[\hat{f}] - \mathbf{R}[f^*] \leq 2\epsilon_{\text{est}} + \epsilon_{\text{app}}, \tag{11}$$

and we got the desired result. □

References

[1] A. Rahimi and B. 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 21*, pages 1313–1320. Curran Associates, Inc., 2009.