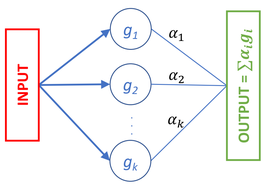
Greedy Convex Ensemble

Tan Nguyen[1], Nan Ye[2], Peter Bartlett[3]
[1]ACEMS, Queensland University of Technology, [2]University of Queensland, [3]UC Berkeley

Introduction

Problem: Learn a convex ensemble of basis models.



Given some set $\mathcal{G}$ of simple basis models

Linear hull, $\text{lin}(\mathcal{G})$, is the set of all possible linear combinations of models in $\mathcal{G}$. $\text{lin}(\mathcal{G})$ of even simple basis models provides universal approximations [1, 2].

Convex hull, $\text{co}(\mathcal{G})$, is the set of all possible convex combinations of models in $\mathcal{G}$. The convex hull is a subset of the linear hull.

Theoretical questions: What is the capacity and generalization property of convex hulls *for general loss functions*? How to efficiently learn from a convex hull?

Empirical questions: How does convex hull fair with competing methods? Is there any advantage for using convex hull?

Contributions

Theory and Algorithms:

- We show that linear hull of very simple basis functions can have unbounded capacity, and is thus prone to overfitting. On the other hand, convex hulls can be seen as a regularized version of linear hulls, and they are still rich but have bounded capacities.
- We extend the generalization bound for convex hulls to a general class of Lipschitz loss functions.
- We propose several greedy algorithms based on iterative functional optimization, and a reparameterization trick to deal with the convex hull constraint.

Empirical contributions:

- We empirically compared the proposed greedy algorithms to find the best.
- We performed a comprehensive empirical comparison of the greedy convex ensemble with competing methods (neural network, boosting, random forest) to show that it is competitive in most cases, while requiring little effort on hyper-parameters tuning.

Definitions

Problem: Given an i.i.d. sample $\mathbf{z} = ((x_1, y_1), \ldots, (x_n, y_n))$ drawn from a distribution $P(X, Y)$ defined on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbf{R}^d \times \mathbf{R}$, learn a function $f \in \text{co}(\mathcal{G})$ to minimize the *expected risk* $R(f) = \mathbb{E}L(Y, f(X))$, where $L(y, \hat{y})$ is the loss that f incurs when predicting y as $\hat{y}$, and the expectation is taken over P.

- $R_n(f) = \mathbb{E}_n L(Y, f(X)) = \frac{1}{n}\sum_{i=1}^n L(y_i, f(x_i))$ denotes the *empirical risk*.
- $\mathcal{G} \subseteq \mathcal{Y}^{\mathcal{X}}$ denotes the *class of basis models/functions*, which are assumed to be *bounded*, i.e. $\mathcal{G} \subseteq \mathcal{Y}^{\mathcal{X}} \subseteq [-B, B]^{\mathcal{X}}$ for some constant $B > 0$.
- $\text{lin}_k(\mathcal{G}) = \{\sum_{i=1}^k \alpha_i g_i : \alpha_i \in \mathbf{R}, g_i \in \mathcal{G}\}$ denotes the set of linear combinations of k functions in $\mathcal{G}$. $\text{lin}(\mathcal{G}) = \cup_{k\geq 1} \text{lin}_k(\mathcal{G})$ denotes the *linear hull* of $\mathcal{G}$.
- $\text{co}_k(\mathcal{G}) = \{\sum_{i=1}^k \alpha_i g_i : \sum_i \alpha_i = 1, \alpha_i \geq 0, g_i \in \mathcal{G}\}$ is the set of convex combinations of k functions in $\mathcal{G}$. $\text{co}(\mathcal{G}) = \cup_{k\geq 1} \text{co}_k(\mathcal{G})$ is the *convex hull* of $\mathcal{G}$.
- Given a class of real-valued functions $\mathcal{F}$, $d_{VC}(\mathcal{F})$ denotes its VC-dimension, $d_P(\mathcal{F})$ denotes its Pseudo-dimension. $\text{bin}(\mathcal{F}) = \{x \mapsto \mathbb{I}(f(x) \geq t) : f \in \mathcal{F}, t \in \mathbf{R}\}$ denotes its thresholded binary version.
- Capacity measures, including Rademacher complexity, see e.g. [3, 4].
- $\mathcal{T} = \{\mathbb{I}(\theta^\top x \geq t) : \theta \in \mathbf{R}^d, t \in \mathbf{R}\}$ is the set of linear threshold functions.

Capacity of Convex Hull - A Regularization Perspective

Proposition 2: *Assume that $d \geq 2$. Then:* **(a)** *$d_{VC}(\text{bin}(\text{lin}_k(\mathcal{T}))) \geq k$, thus $d_P(\text{lin}_k(\mathcal{T})) \geq k$, and $d_P(\text{lin}(\mathcal{T})) = \infty$. In addition, the Rademacher complexity of $\text{lin}(\mathcal{T})$ is infinite.* **(b)** *$d_{VC}(\text{bin}(\text{co}_{k+1}(\mathcal{T}))) \geq k$, thus $d_P(\text{co}_{k+1}(\mathcal{T})) \geq k$, and $d_P(\text{co}(\mathcal{T})) = \infty$, but the Rademacher complexity of $\text{co}(\mathcal{T})$ is finite.*

Implications: *Linear hulls* of even simple basis models have *unbounded capacity*, thus are prone to *overfitting*. In contrast, *convex hulls* are rich (unbounded pseudo-dimension) but have *bounded capacity* (finite Rademacher complexity). Thus, convex hulls can be seen as a *regularized version of the linear hulls* with weights of basis model constrained to be inside a simplex.

Generalization Error Bound of Convex Hulls

Assume that $d_P(\mathcal{G}) = p < \infty$, and $L(y, f(x)) = \phi(f(x) - y)$ for a c_ϕ-Lipschitz non-negative function ϕ satisfying $\phi(0) = 0$. Let $c = 2c_\phi B\left(\sqrt{2\ln(1/\delta)} + D\sqrt{p} + 2\right)$, where D is an absolute constant. Then:

Theorem 2: *Let $f^*(x) = \min_{y\in\mathcal{Y}} \mathbb{E}[L(y, Y)|X = x]$ be the Bayes optimal function. With probability at least $1-\delta$, for all $f \in \text{co}(\mathcal{G})$, $R(f) - R(f^*) \leq R_n(f) - R_n(f^*) + \frac{c}{\sqrt{n}}$.*

$\implies$ Minimizing the empirical risk $R_n(f)$ over the convex hull $\text{co}(\mathcal{G})$ results in minimizing the bound of the expected risk $R(f)$.

Theorem 3: *Let $\hat{f} = \arg\min_{f\in\text{co}(\mathcal{G})} R_n(f)$, and $h^* = \arg\min_{f\in\text{co}(\mathcal{G})} R(f)$, then with probability at least $1 - \delta$, $R(\hat{f}) \leq R(h^*) + \frac{c}{\sqrt{n}}$.*

$\implies$ In $\text{co}(\mathcal{G})$, empirical risk minimizer $\hat{f}$ converges to the expected risk minimizer h^* at the rate of $O(1/\sqrt{n})$ as number of samples n increases.

Note: Similar generalization bound has been proven for quadratic loss [5]. We prove a more general bound for a class of Lipschitz losses that includes the quadratic loss as a special case. Proofs are available at `github.com/tan1889/gce`.

Greedy Algorithms to Learn from Convex Hulls

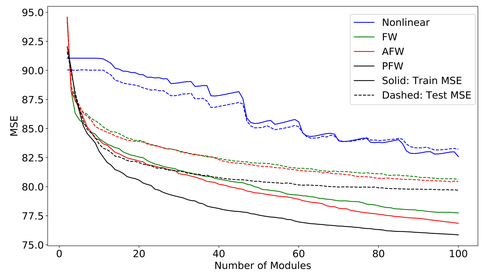
Goal: Find empirical risk minimizer $\hat{f} = \arg\min_{f\in\text{co}(\mathcal{G})} R_n(f)$, where $\mathcal{G} = \{g_\theta : \theta \in \mathbf{R}^p\}$ is the set of basis models g_θ parameterized by $\theta \in \mathbf{R}^p$.

Greedy scheme: Start with some $f_0 \in \mathcal{G}$. At iteration t, we choose appropriate $\alpha_t \in [0, 1]$ and $g_t \in \mathcal{G}$ for the new convex combination $f_{t+1} = (1 - \alpha_t)f_t + \alpha_t g_t$. Repeat up to T iterations, or *stop early* if improvement hits plateau.

Non-linear greedy: Choose g_t and α_t jointly to maximize the decrease in empirical risk, i.e. $\theta_t, \alpha_t \leftarrow \arg\min_{\theta\in\mathbf{R}^p, \alpha\in[0,1]} R_n((1 - \alpha)f_{t-1} + \alpha g_\theta)$. If this step is solved with an error of $O(1/t^2)$ then f_t converges at rate $O(1/t)$ [6].

Frank-Wolfe (FW): Choose g_t that aligns the most with the negative functional gradient, i.e. $g_t = \arg\min_{\theta\in\mathbf{R}^p} \langle DR_n(f_{t-1}), g_\theta\rangle$. Choose $\alpha_t = \frac{1}{t+1}$ or use line search. FW converges at rate $O(1/t)$ [7]. **Away-step Frank-Wolfe (AFW), Pair-wise Frank-Wolfe (PFW)** are variants of FW that *converges linearly* if the solution is in the interior of the convex hull [8].

Empirical Comparison of GCE Algorithms



We empirically compared the above GCE algorithms and found that PFW is the best in most cases. The graph shows the training and testing performance of these algorithms on the MSD dataset.

Empirical Comparison of GCE with Competing Methods

Empirical comparison of GCE with Boosting, Random Forest, and Neural Network on a variety of datasets shows that GCE is competitive in the majority of cases.

Datasets	#Samples	GCE	XGBoost	RForest	NN	ConvNet
diabetes	442	**42.706**	46.569	49.519	43.283	44.703
boston	506	**2.165**	2.271	2.705	2.217	2.232
ca_housing	20,640	0.435	**0.393**	0.416	0.440	0.437
msd	515,345	**6.084**	6.291	6.462	6.186	7.610
iris	150	**0.00**	6.67	6.67	3.33	10.00
wine	178	**0.00**	2.78	2.78	**0.0**	**0.0**
breast_cancer	569	**3.51**	4.39	8.77	3.51	4.39
digits	1,797	2.78	3.06	**2.50**	3.33	3.06
cifar10_f	60,000	**4.86**	5.40	5.16	5.00	4.92
mnist	70,000	1.22	1.66	2.32	1.24	**1.11**
covertype	581,012	26.70	26.39	27.73	26.89	**26.56**
kddcup99	4,898,431	**0.01**	**0.01**	**0.01**	**0.01**	**0.01**

References

[1] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. IEEE ToIT, 1993.
[2] Y. Makovoz. Random approximants and neural networks. Journal of Approximation Theory, 1996.
[3] M. Anthony and P. L. Bartlett. Neural network learning: Theoretical foundations. Cambridge University Press, 2009.
[4] S. Mendelson. A few notes on statistical learning theory. Advanced lectures on machine learning, p1-40. Springer, 2003.
[5] S. Mannor, R. Meir, T. Zhang. Greedy algorithms for classification - consistency, convergence, adaptivity. JMLR, 2003.
[6] T. Zhang. Sequential greedy approximation for certain convex optimization problems. IEEE ToIT, 2003.
[7] M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. ICML, 2013.
[8] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. NIPS, 2015.