Diversity and Generalization in Neural Network Ensembles

Luis A. Ortega, Rafael Cabañas and Andrés R. Masegosa

March 11, 2022

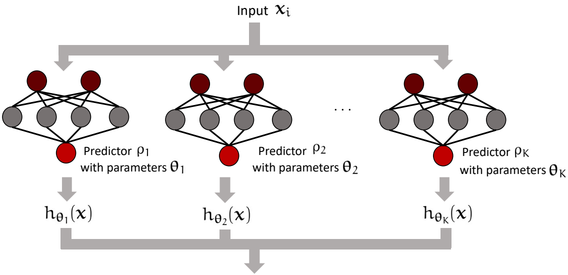
AISTATS 2022

Introduction

- An ensemble is **diverse** if the predictions of individual models do not coincide on all the samples.
- Diversity is known to improve ensemble performance, but **why?**.
- We employ a PAC-Bayesian analysis to answer these questions:
  1) How to measure the diversity of an ensemble?.
  2) How is diversity related to the generalization error?.
  3) How can diversity be promoted by ensemble learning algorithms?.

Basics on NNs ensembles

An ensemble trained with $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ is the combination of different predictors.



Averaging methods. Model averaging and Majority vote.

Loss functions. Squared loss, Cross-entropy loss and Zero-one loss.

Diversity and Generalization

Decomposing the loss using upper bounds

General Upper-bound for all the ensembles considered in this work:

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's} \\ \text{Expected Error}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Individual Models'} \\ \text{Expected Error}}} - \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's} \\ \text{Diversity}}} \right)$$

where $\alpha = 4$ for the $0/1$-loss, otherwise, $\alpha = 1$.

The diversity term depends on the considered **loss function**:

$$\mathbb{D}(\rho) = \mathbb{E}_\nu\Big[\mathbb{V}_\rho\left(f(y, \boldsymbol{x}; \boldsymbol{\theta})\right)\Big].$$

E.g. $\mathbb{D}_{sq}(\rho) = \mathbb{E}_\nu\Big[\mathbb{V}_\rho(h_R(\boldsymbol{x}; \boldsymbol{\theta}))\Big].$

Diversity and Generalization

Decomposing the loss using upper bounds

General Upper-bound for all the ensembles considered in this work:

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's}\\\text{Expected Error}}} \;\leq\; \alpha \left(\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Individual Models'}\\\text{Expected Error}}} \;-\; \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's}\\\text{Diversity}}} \right)$$

where $\alpha = 4$ for the 0/1-loss, otherwise, $\alpha = 1$.

An ensemble ρ outperforms a single model $\boldsymbol{\theta}^\star$ when the diversity of the ensemble is large enough:

$$\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})] - \frac{1}{\alpha} L(\boldsymbol{\theta}^\star)}_{\substack{\text{Single Model's}\\\text{Error Gap}}} < \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's}\\\text{Diversity}}} \implies \underbrace{L(\rho)}_{\substack{\text{Ensemble's}\\\text{Expected Error}}} < \underbrace{L(\boldsymbol{\theta}^\star)}_{\substack{\text{Single Model's}\\\text{Expected Error}}}.$$

Diversity and Generalization

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's}\\\text{Expected Error}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[L(\boldsymbol{\theta})]}_{\substack{\text{Individual Models'}\\\text{Expected Error}}} - \underbrace{\mathbb{D}(\rho)}_{\substack{\text{Ensemble's}\\\text{Diversity}}} \right)$$

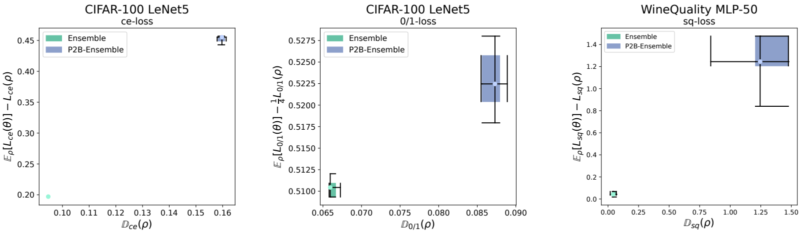
Pac-Bayesian Bound

With probability at least $1-\xi$ over draws of training data $D \sim \nu^n(y, \boldsymbol{x})$, for all distribution ρ over $\boldsymbol{\Theta}$, simultaneously,

$$\underbrace{L(\rho)}_{\substack{\text{Ensemble's}\\\text{Expected Error}}} \leq \alpha \left(\underbrace{\mathbb{E}_\rho[\hat{L}(\boldsymbol{\theta}, D)]}_{\substack{\text{Empirical}\\\text{Averaged Error}}} - \underbrace{\hat{\mathbb{D}}(\rho, D)}_{\substack{\text{Ensemble's}\\\text{Empirical Diversity}}} + \underbrace{\frac{2KL(\rho \mid \pi) + \epsilon}{n}}_{\text{Regularization}} \right).$$

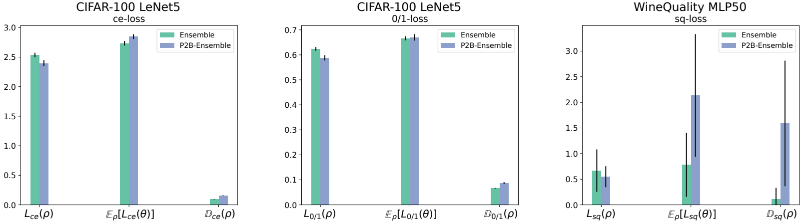
Empirical evaluation

Relation of the diversity and the gap between the losses of the individual models and the one of the ensemble.



Empirical evaluation

Ensemble error, average individual models errors and ensemble diversity.



Poster Session

Tue 29 Mar

10 a.m. CEST — 11:30 a.m. CEST