Stage de Master 2019

Apprentissage causal: Approche adversariale

SUPERVISOR: Michèle Sebag, sebag@lri.fr
LAB: LRI − CNRS − INRIA − Paris-Sud, U. Paris-Saclay

RÉSUMÉ :
Le contexte est celui de l'apprentissage supervisé. Etant donné une base d'entrainement iid selon la distribution $P(x, y)$, on apprend classiquement un modèle h qui minimise la perte $\mathcal{L}(h(x), y)$ en espérance selon $P(x, y)$. Mais dans le contexte de l'apprentissage causal (ainsi que pour l'adaptation de domaine, ou pour l'apprentissage dans des domaines critiques), on souhaite des garanties plus fortes sur la perte du modèle appris [1]. La robustesse est le goulet d'étranglement de l'IA nouvelle.

Le but du stage est de proposer une formalisation de l'apprentissage robuste par rapport à la distribution des données d'entrainement, et une solution algorithmique efficace. Ce stage peut déboucher sur une thèse.

Objectifs du stage
Idéalement, l'apprentissage causal aimerait obtenir des modèles de perte minimale en norme infinie:

$$\text{Trouver } \hat{h} = \arg\min_h \max_x \mathcal{L}(h(x), y) = ||\mathcal{L}(h(x), y)||_\infty$$

La formulation ci-dessus n'est pas réaliste : si les données sont bruitées (ce qui est toujours le cas dans les données réelles), l'erreur L_∞ même d'un très bon modèle peut être arbitrairement grande.

Approche proposée
Cette approche s'inspire des mécanismes adversariaux proposés par Goodfellow et al. [2], et Ganin et al [3]. Formellement, l'approache fait intervenir deux agents. Le premier (agent échantillonnant, AE) agit sur la distribution des données d'entrainement; le second (agent apprenant, AA) apprend un modèle sur les données d'entrainement, dont la qualité est évaluée sur les données de validation (disjointes des données d'entrainement). Le mécanisme proposé pour obtenir et minimiser une approximation de la perte L_∞ correspond à un problème max min: chercher le maximum (pris sur les AE) du minimum (pris sur les AA) de la perte en validation du modele.

Tel quel, ce mécanisme donne trop d'avantage à l'AE (e.g. il suffirait d'oublier toutes les données d'une région de l'espace pour obtenir une perte très mauvaise). On rajoute donc une contrainte supplémentaire, imposant que l'agent échantillonnant ne puisse pas trop s'éloigner de la distribution marginale $P(x)$ des données d'apprentissage. Si $Q(x)$ est la distribution imposée par AE, le critère devient:

$$\text{Trouver } min_h \; max_Q \; \mathbb{E}_Q[\mathcal{L}(h(x), y)] \text{ s.t. } KL(P||Q) < C$$

NB: on pourra procéder en considérant que l'échantillonneur définit en fait une pondération (importance sampling) sur les données.

Ce stage demande de très bonnes capacités mathématiques et informatiques (programmation NN en partant de librairies existantes).

Références

[1] Jonas Peters, Dominik Janzing and Bernhard Schölkopf, Elements of Causal Inference, MIT 2017.
[2] Ian J. Goodfellow: NIPS 2016 Tutorial: Generative Adversarial Networks.
[3] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, Victor S. Lempitsky: Domain-Adversarial Training of Neural Networks. JMLR 2017.