Симпозијум МАТЕМАТИКА И ПРИМЕНЕ, Математички факултет, Универзитет у Београду, 2022 ,Vol. XII (1)

Ублажавање пристрасности неуронских мрежа обртањем градијента

Лазар Васовић

Математички факултет, Универзитет у Београду, Студентски трг 16, Београд
e-mail: pd212006@alas.matf.bg.ac.rs

Апстракт. Робусност модела машинског учења огледа се у приближно једнаком успеху над разним подацима. Уколико то није случај, модел је пристрастан. Посебан проблем са пристрасношћу имају неуронске мреже, будући да су врло флексибилне и (пре)прилагодљиве подацима над којима су обучене. Као решење овог проблема, примењују се бројне опште технике регуларизације, попут случајног изостављања неурона. С друге стране, постоје специфични приступи, који могу отклонити неке конкретне изворе пристрасности. У раду је представљено обртање градијента, као пример таквог приступа. Ова техника одликује се гранањем неуронске мреже, при чему постоји заједнички део, који се затим дели на две гране – главну и супарничку. Главна грана решава конкретан проблем (нпр. класификација), док супарничка открива евентуални извор пристрасности (углавном домен инстанце, нпр. пол у случају људи). Приликом ажурирања параметара модела, заснованог на градијентној минимизацији грешке, градијент се обрће (негира) на заједничком делу супарничке гране. Овиме се ефективно отклања извор пристрасности, па је резултујући модел (представљен главном граном) робуснији. Као практични допринос рада, обртање градијента је примењено на познати скуп рецензија са сајта Амазон, у циљу предвиђања да ли је оцена производа позитивна или негативна. Добијени резултати у свим експериментима који укључују супарничку грану бољи су од оних који је не укључују.

Кључне речи: пристрасност; неуронске мреже; обртање градијента; супарничко учење.