IP-feladatok futásidejének becslése

Adatbányászat projekt

Becsó Gergely, Dankó Dorottya

2022. december

Az adathalmaz

- 5000 db IP-feladat: $A\mathbf{x} \leq \mathbf{b}$, $\max(\mathbf{cx})$
  - $A \in \{0,1\}^{100\times 200}$
  - $\mathbf{b} = \mathbf{1}$
  - $\mathbf{c} \in \mathbb{Z}^{200}$, c elemei 0 és 100 közé esnek
- SCIP7 IP-solverrel
- futásidők, felhasznált LP-feladatok száma

	Átlag	Szórásnégyzet
Futásidő	0.139839	0.021164
LP-k száma	12.6888	238.7407

Előfeldolgozás

- .lp formátumú fájlok beolvasása
- pandas dataframe: az A mátrixok és a c vektorok az elemek (ezt a változatot használtuk a neurális hálókhoz)
- „kilapítás”: 5000×20205 méretű pandas dataframe létrehozása, minden mátrix- és vektorelemhez tartozik egy oszlop (ezt használtuk a klasszikus regressziókhoz)
- egyedi pyTorch DataSet létrehozása

Klasszikus regressziók

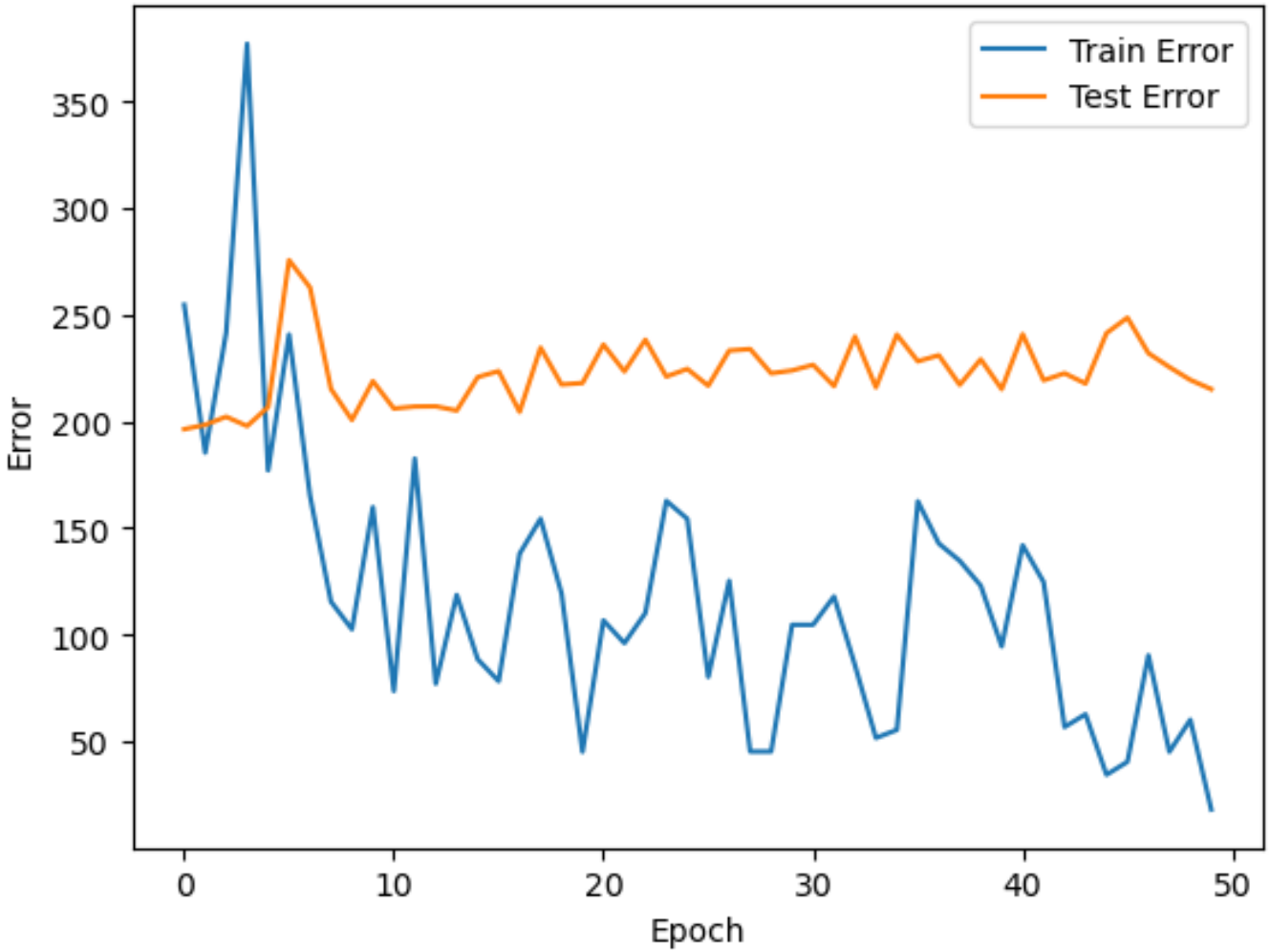
- tanuló adat: első 4500 sor, teszt: utolsó 500 sor
- Lineáris regresszió: Scikit Learn LinearModel LinearRegression
- Extreme Gradient Boosting: xgboost XGBRegressor modell

	MSE (futásidő)	MAE (futásidő)	MSE (LP-k)	MAE (LP-k)
LinearRegression	0.02338	0.12330	271.9823	13.2126
XGBRegressor	0.02059	0.11538	241.16588	12.6485

Egyszerű hálók

- tanuló adat: első 4500 sor, teszt: utolsó 500 sor
- sűrű rétegekből álló architektóra, 20000, 512 és 1 méretekkel, ReLU aktivációs függvény
- variációk a belső réteg méretével, dropout rétegek és plusz belső rétegek segítségével
- Adam optmizer, weight decay használata
- túltanulás az első pillanattól kezdve

Teszt- és tanulási hibák



1. ábra. A deep dense dropout háló teszt- és tanulási hibái

Összetett háló

- tanuló adat: első 4500 sor, teszt: utolsó 500 sor
- a feladat feltételleit LSTM háló segítségével beágyazzuk $\mathbb{R}^{64}$-be
- attention réteg segítségével (hibás kódolás)
- két sűrű rétegből álló fej, hogy az elkódolt vektorokból egy számértéket kapjunk
- Adam optimizer, a korábbi eredményekhez hasonló túltanulás

Konklúzió

A kipróbált modellek egyike sem tudott lényegi tudást kinyerni az adatból. Lehetséges, hogy az adatgenerálásnál csúszott be hiba, esetleg a címkék keveredtek össze, vagy nincs lényegi összefüggés a vizsgált IP-feladatok alakja és a futásidő között. További kutatás szükséges.