Dimenziócsökkentés szubmoduláris kiválasztással

Készítette: Bartalis Dávid

Témavezetők:
Bérczi-Kovács Erika
ELTE, Operációkutatási Tanszék
Béres Ferenc
SZTAKI, Informatikai Kutatólaboratórium

Budapest, 2021

A félév során elvégzett feladatok

- Az **Apricot** Python csomaghoz, illetve a dimenziócsökkentési eljárásokhoz kapcsolódó szakirodalom áttekintése.
- A **Comet ML**, mint mérési felület megismerése.
- Az Apricot csomagban implementált függvények paraméterezésének vizsgálata, legjobb paraméter megkeresése.
- Az implementált mohó algoritmusok tesztelése (teljesítményt és futásidőt tekintve).
- Az Apricot és a **PCA** (Principal Component Analysis) összehasonlítása, mérések végzése.

Feldolgozott cikkek

- Xiao-Zhong Zhu, William Zhu, Xin-Nan Fan:
  Rough set methods in feature selection via submodular function
- Girija Attigeri, Manohara Pai M. M., Radhika M. Pai:
  Feature Selection Using Submodular Approach for Financial Big Data
- Yuzong Liu, Kai Wei, Katrin Kirchhoff, Yisong Song, Jeff Bilmes:
  Submodular Feature Selection for High-Dimensional Acoustic Score Space
- K J Joseph , Vamshi Teja R , Krishnakant Singh , Vineeth N Balasubramanian:
  Submodular Batch Selection for Training Deep Neural Networks
- Maxwell W. Libbrecht, Jeffrey A. Bilmes, William Stafford Noble:
  Eliminating redundancy among protein sequences using submodular optimization

Apricot - Szubmodularitás

Cél: Reprezentatív részhalmaz kiválasztása.

Felhasználás: Tanítási halmaz redukálása, tanulási folyamat felgyorsítása.

Módszer: Szubmoduláris kiválasztás.

Definíció: Egy $\mathcal{F} : 2^V \to \mathbb{R}$ halmazfüggvény **szubmoduláris**, ha $\forall B \subseteq A \subseteq V$ halmazokra és $x \in \overline{\mathrm{A}}$ esetén

$$\mathcal{F}(A \cup x) - \mathcal{F}(A) \leq \mathcal{F}(B \cup x) - \mathcal{F}(B).$$

Használt függvények

Feature-based / Tulajdonság alapú függvény:

$$\mathcal{F}(X) = \sum_{d=1}^{D} w_d \phi \left(\sum_{x \in X} m_d(x) \right)$$

Facility location / Szolgáltató elhelyezési függvény:

$$\mathcal{F}(X) = \sum_{v \in V} \max_{x \in X} \delta(x, v)$$

Max Coverage / Maximális fedés függvény:

$$\mathcal{F}(X) = \sum_{i=1}^{d} \left(\left(\sum_{x \in X} x_i \right) > 0 \right)$$

Használt algoritmusok I.

Naive greedy: Minden iterációban sorra veszi a még nem kiválasztott összes elemet és kiszámolja, hogy melyiknek mennyi a hozzáadott értéke. A legnagyobb hozzáadott értékű elemet fogja kiválasztani.

Lazy greedy: Itt egy maximum prioritású sorban tároljuk a még nem választott elemeket a legutoljára kiszámolt hozzáadott érték szerint.

Two-stage greedy: Az első k lépésben a naív módszert használjuk, majd átváltunk a lazy algoritmusra.

Használt algoritmusok II.

Stochastic greedy : Minden iterációjában véletlenül választ egy részhalmazt, amiből a következő elemet választja.

Sample greedy: Az algoritmus elején egy véletlen mintavételezés történik.

Approximate lazy: Nem feltétlenül azt az elemet vesszük hozzá a már kiválasztott részhalmazhoz, ami a legnagyobb haszonnal jár, hanem ami "közel legjobb haszonnal jár".

PCA - Principal Component Analysis

Főkomponens analízis.

Lényeges lépések:
M mátrix sorai az adatpontok $\rightarrow$ $M^T M$ kovarianciamátrix kiszámolása $\mathcal{O}(nD \min(n, D)) \rightarrow$ $M^T M$ sajátvektorainak és sajátértékeinek számítása $\mathcal{O}(D^3)$.

Futásidő:
Mindezt összevetve az algoritmus futási ideje $\mathcal{O}(nD \min(n, D) + D^3)$.

Elért eredmények

- NLP (Natural Language Processing) feladat
- Feladat: egy adott tweet valóban katasztrófáról szól-e
- Tanítási halmaz mérete: (5264, 10000)

	id	text	target
3806	5408	Former Township fire truck being used in Phili...	0
3444	4922	The Dress Memes Have Officially Exploded On Th...	0
3443	4920	Well as I was chaning an iPad screen it fuckin...	0
6219	8875	So does Austin smoke too since he agreed to th...	0
3440	4917	Im Dead!!! My two Loves in 1 photo! My Heart e...	0
...	...	...	...
3663	5213	@Truly_Stings Yo Dm me	1
3660	5210	Driver fatalities down on Irish roads but pede...	1
3659	5209	Message boards will display updated traffic fa...	1
3673	5228	Kosciusko police investigating pedestrian fata...	1
7612	10873	The Latest: More Homes Razed by Northern Calif...	1

Előző félév eredménye

Dimenziócsökkentéshez jól működött a módszer.
Tanításhoz használt modellek: Logisztikus Regresszió
Kiértékeléshez használt metrikák:

- Accuracy
- Precision
- Recall
- Roc Auc

Legjobb eredmény a **feature based** függvénnyel.

A mohó algoritmusok összehasonlítása - futásidő

Optimizer \ Size	1 %	5%	10 %	25%	50%
approximate-lazy	9.87	38.29	85.97	318.35	838.84
lazy	**5.85**	**8.65**	**24.38**	**166.21**	**670.32**
naive	8.79	21.41	49.76	216.48	783.08
sample	12.15	47.94	96.67	257.53	870.33
stochastic	8.07	18.83	43.26	217.42	785.12
two-stage	6.71	24.62	50.72	238.43	827.12

Table: A vizsgált mohó algoritmusok összehasonlítása futásidőt tekintve.

A mohó algoritmusok összehasonlítása - teljesítmény

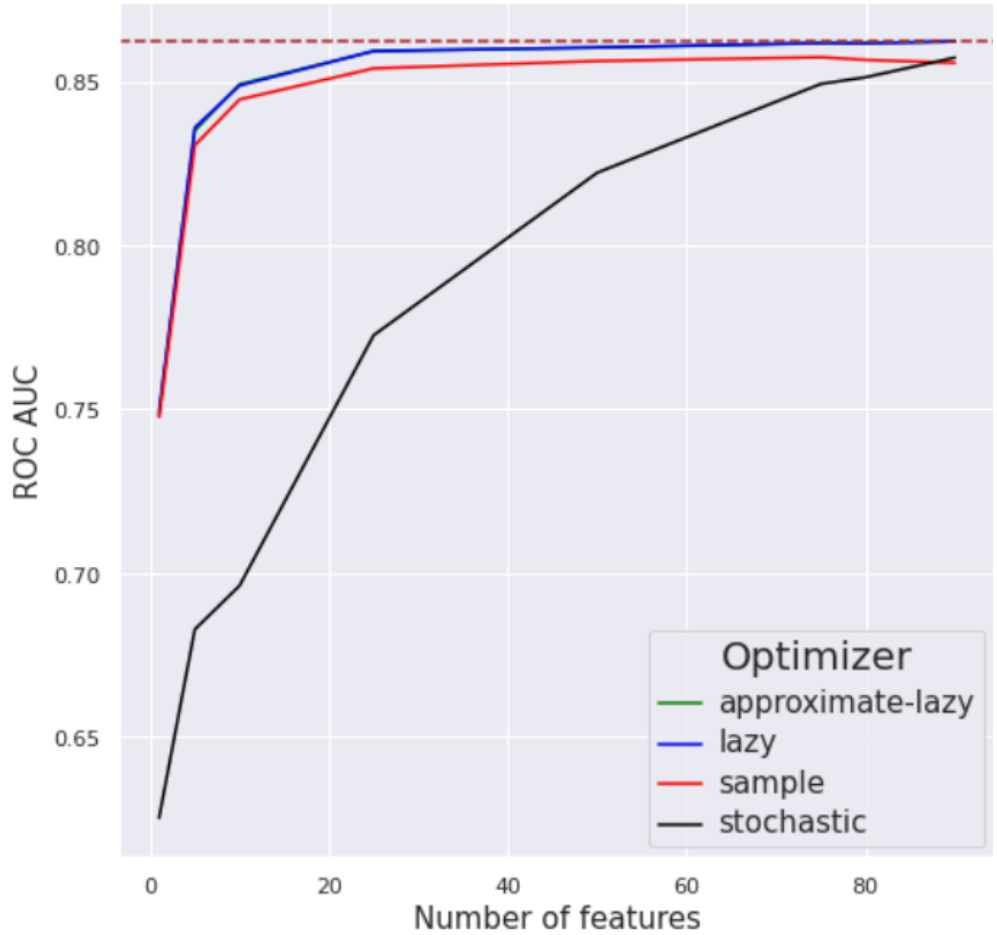

Figure: A vizsgált mohó algoritmusok összehasonlítása ROC-AUC értéket tekintve. A szaggatott vonal a dimenziócsökkentés nélkül tanított logisztikus regresszióval elért eredményt mutatja.

A feature-based függvény legjobb paraméterezése

Size	Best parameters	
%	Concave Function	Optimizer
1 %	"log"	"lazy"
5 %	"log"	"approximate-lazy"
10 %	"sqrt"	"approximate-lazy"
25 %	"log"	"approximate-lazy"
50 %	"log"	"lazy"

Table: A feature-based függvény legjobb paraméterezése a különböző méretű redukciók során.

Az Apricot összehasonlítása a PCA módszerrel

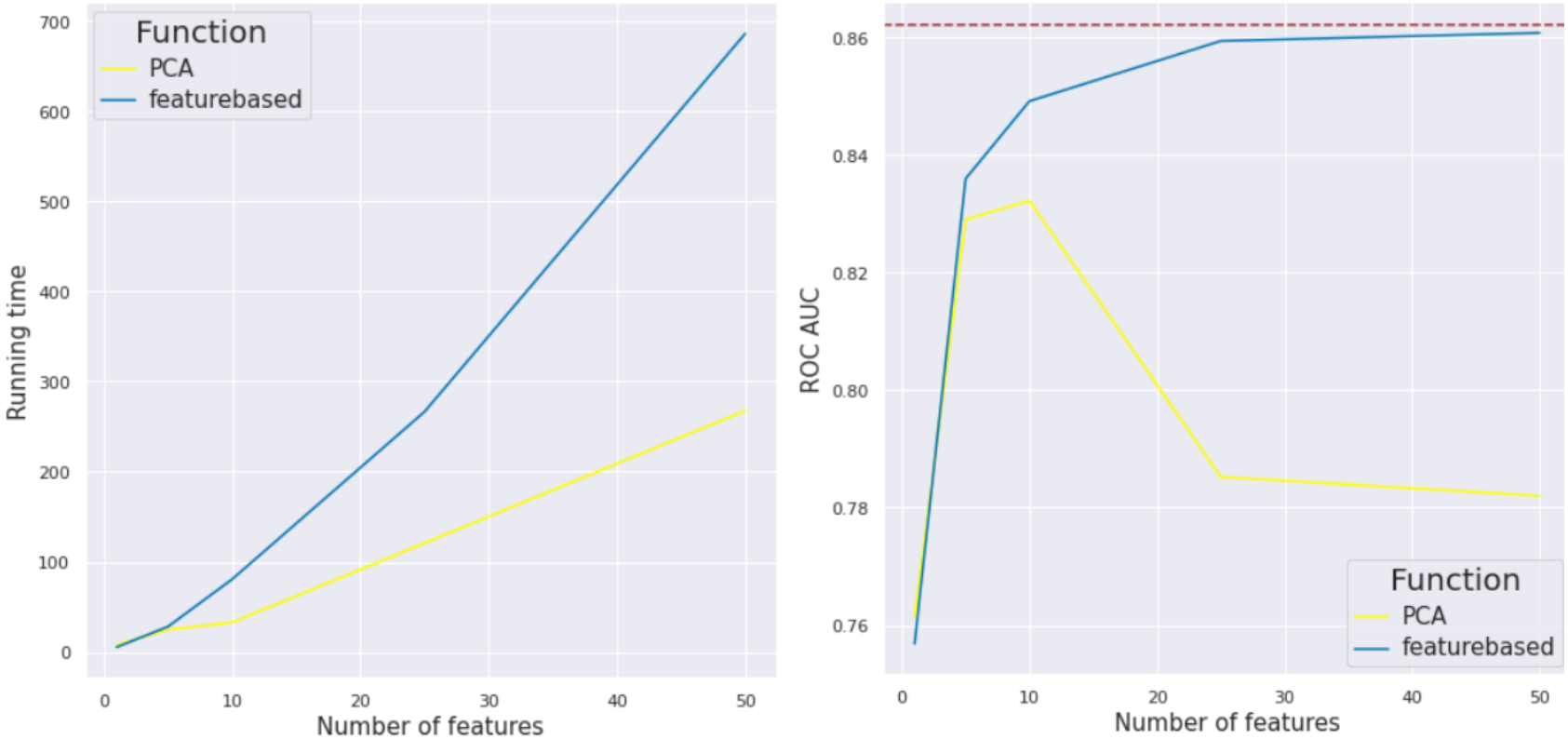

Figure: Az Apricot és a PCA módszer összehasonlítása futásidőt (bal), illetve teljesítményt (jobb) tekintve. A jobb oldali ábrán a szaggatott vonal a dimenziócsökkentés nélkül tanított logisztikus regresszióval elért eredményt szemlélteti.

Összegzés, tervek

Összegzés:
Van létjogosultsága a szubmoduláris függvényekkel való adathalmaz redukciónak. Főleg a dimenziócsökkentés területén.

Továbbiak:
Tervezzük folytatni a munkát. Fő célok:

- Egy még nem implementált szubmoduláris függvény írása a csomagba. Jelenleg a legjobb jelölt egy entrópia megközelítést használó függvény.
- További példák keresése, ahol a szubmoduláris megközelítés jól működik.

Köszönöm a figyelmet!