Adattömörítés szubmoduláris kiválasztással

Készítette: Bartalis Dávid

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

Budapest, 2020

Tartalomjegyzék

A félév során elvégzett feladatok

- Az Apricot csomag működésének vizsgálata
- Adatbányászati kurzusok elvégzése (ML, Pandas)
- A Kaggle, illetve Kaggle notebookok használatának megismerése
- Egy Kaggle adathalmaz elemzése, vizsgálata
- Az Apricot módszer kipróbálása az adathalmazon

Apricot

Az Apricot egy Python csomag, aminek segítségével hatalmas adathalmazból ki lehet választani egy olyan részhalmazt, ami az egész adatsokaságot reprezentálja.

Felhasználás: Tanítási halmaz redukálása, tanulási folyamat felgyorsítása.

Módszer: szubmoduláris kiválasztás.

Szubmodularitás

Definíció

Egy $\mathcal{F} : 2^V \to \mathbb{R}$ V alaphalmaz részhalmazain értelmezett halmazfüggvényt szubmodulárisnak, vagy teljesen szubmodulárisnak nevezünk, ha $\forall X, Y \subseteq \mathrm{V}$ halmazpárra teljesül a következő egyenlőtlenség:

$$\mathcal{F}(X) + \mathcal{F}(Y) \geq \mathcal{F}(X \cap Y) + \mathcal{F}(X \cup Y)$$

Állítás

Egy $\mathcal{F} : 2^V \to \mathbb{R}$ V alaphalmaz részhalmazain értelmezett halmazfüggvény pontosan akkor szubmoduláris, ha $\forall B \subseteq A \subseteq V$ halmazokra és $x \in \overline{\mathrm{A}}$ esetén igaz, hogy

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

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:

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

Kaggle, az adathalmaz

- Miért pont a Kaggle?
- Az adathalmaz: Titanic: Machine Learning from Disaster

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

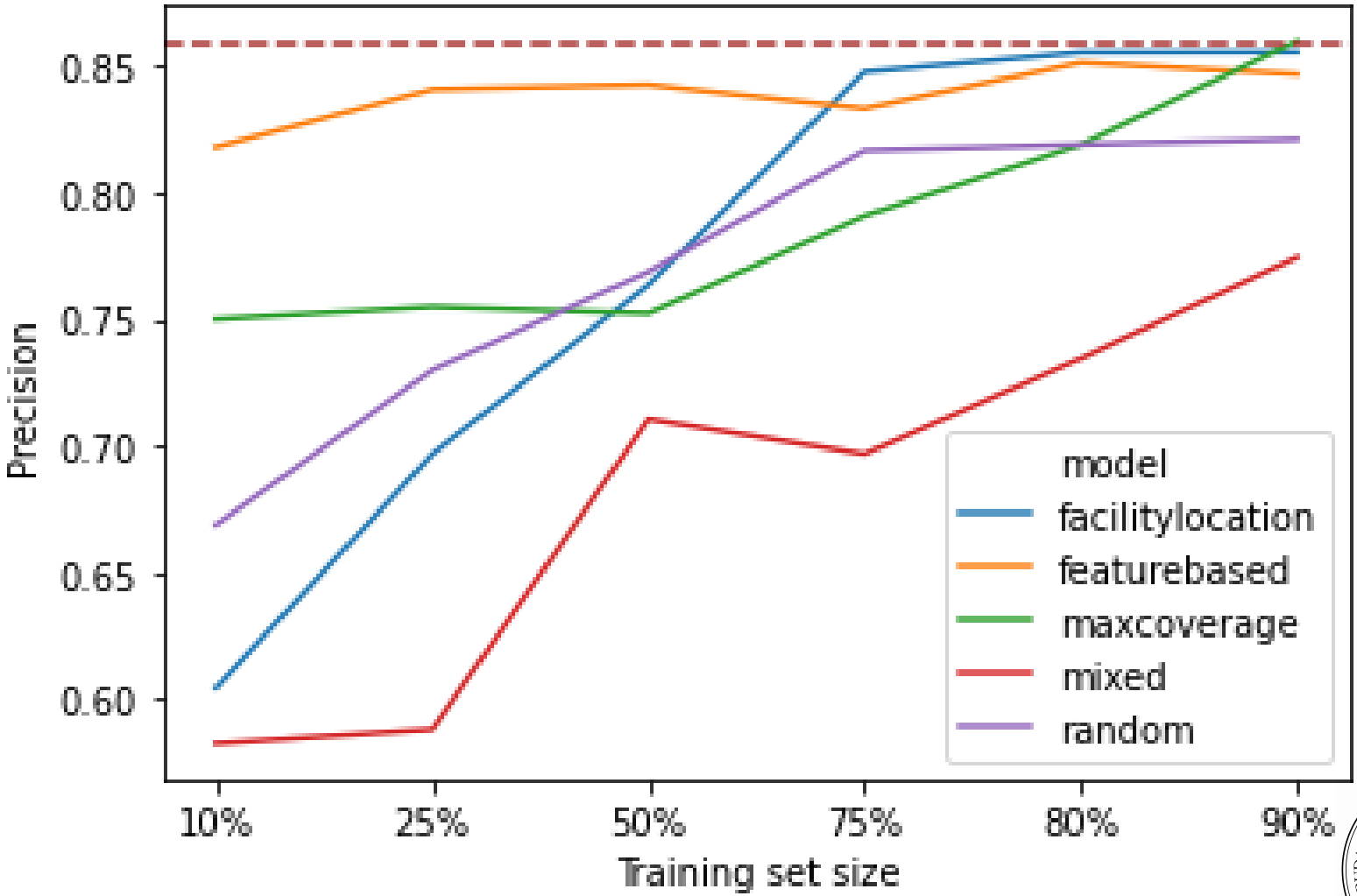
Modell

Feladat: Egy adott emberről megjósolni, hogy túlélte-e a katasztrófát.
Legjobb eredményt elérő modell: Gradient Boosting Classification
Kiértékeléshez használt metrikák:

- Accuracy
- Precision
- Recall
- Roc Auc

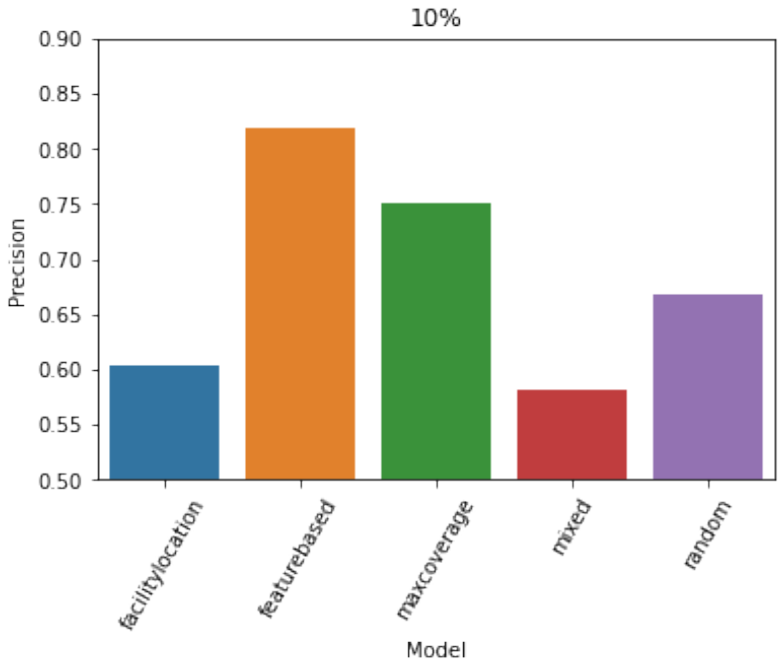
Cél: a tanítási adathalmaz sorainak redukálása az Apricot-val (10%-ra, 25%-ra, 50%-ra, 75%-ra, 80%-ra, 90%-ra)

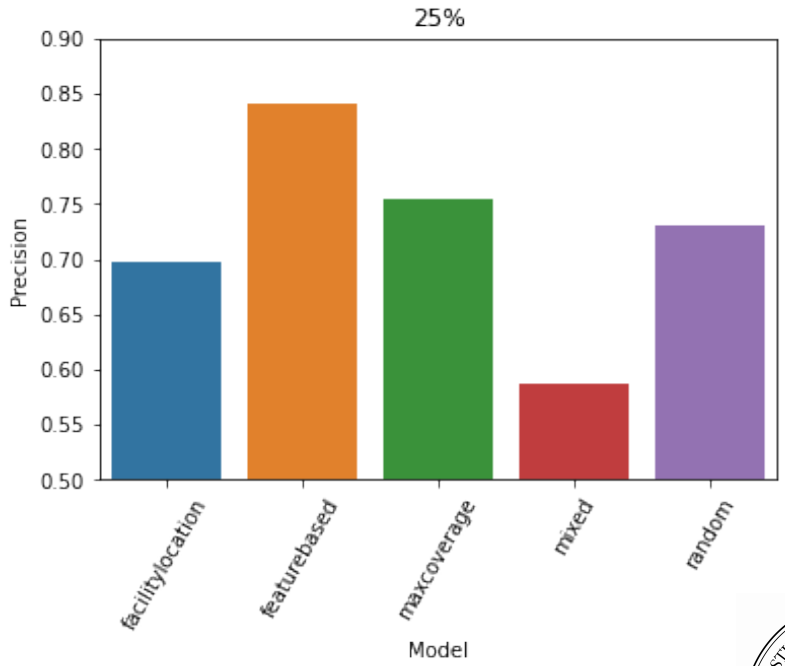
Elért eredmények



Elért eredmények

A szubmoduláris kiválasztás ereje a legjobban az adathalmaz $10 - 25\%$-ára való csökkentésekor látszik. Itt például a precision értékeket nézve a random módszer rendre gyengébb eredményt mutat, mint a feature-based.





Összegzés, tervek

Megállapítás:

- Az Apricot nagyobb, redundánsabb adathalmazokon valószínűleg jobban működik.

Tervek:

- További adathalmazokon is kipróbálni az Apricot módszert.
- Hasznos lenne, ha észre tudnánk venni valami szabályszerűséget abban, hogy mely feladatokhoz, illetve kiértékelési metrikákhoz melyik szubmoduláris függvény használata a legjobb.
- Emellett egy érdekes feladat lenne még egy saját függvény implementálása is.

Köszönöm a figyelmet!