Gépi 1

Definíció: Ha adott egy konkrét T feladat és P teljesítménymetrika, akkor gépi tanulásról beszélünk, ha a rendszer egyre több E tapasztalat/megfigyelés begyűjtése esetén egyre jobban tudja megoldani a T feladatot a P -ben mérve. (Mitchell '97 definíciója)

1. Felügyelet nélküli gépi tanulás

A cél hogy az adatok közt mintázatokat összefüggéseket keressünk.

(pl.: találjunk olyan ügyfél csoportokat akik hasonlóan viselkednek)

2. Felügyelt gépi tanulás

A felügyelt gépi tanulásnál a cél, hogy a modellt korábban nem látott példákra is helyesen működjön. (pl.: ügyfélszolgálat email továbbítás)

A felügyelt gépi tanulásnál megkülönböztetünk 2 féle feladattípust:

Osztályozás

Az osztályozásnál adottak az osztályok és a feladat, hogy ezeknek az osztályoknak valamelyikébe besoroljuk. A célváltozó diszkrét ebben az esetben. Továbbá megkülönböztetünk binary classification illetve multyclassificationt.

Regresszió

A regresszió esetén a célváltozó folytonos érték és a feladtunk, hogy megjósoljuk, hogy az egyedhez melyik érték tartozhat.

3. A gépi tanulás fejlesztési ciklus lépései

1. Adatgyűjtés
2. Előfeldolgozás
3. Jellemző kinyerés
4. Módszer választás
5. Tanítás
6. Kiértékelés

Jellemzőkinyerés szövegből

Ahhoz hogy tanuljunk az egyedeket jellemzőkkel kell leírnunk. Szózsák modellt alkalmazzuk, amely során egy szótárt alkotunk, bekerül az összes szó. Ez a szótár fogja tartalmazni, hogy az egyes szavak hányszor fordultak elő a szövegben. Unigram. Bigram: szó+utána, Trigram szó+előtte+utána.

4. Kiértékelési módszerek

Felügyelt gépi tanulás

baseline: osztályozó -> most frequement class, regresszió -> átlag, medián

Ahhoz hogy a felügyelt gépi tanulási megoldásnak legyen általánosítási készsége, hogy a modell korábban nem látott példákra is pontos eredményt predikáljon. Két részre bontjuk a modellt tanító és kiértékelő adatbázisra. A tanulás során csak a tanító adatbázist láthatja a modell, ezzel szimuláljuk a kiértékelő adatbázis „nem látott példák” lesznek.

A legismertebb tanító/kiértékelő módszer a k-szoros kresztvalidáció. Itt a rendelkezésre álló adatbázist k egyenlő méretű részre bontjuk, mindegyiken tanítunk és kiértékelünk. Minden kísérletben az egyik részhalmaz lesz a kiértékelő adatbázis a többi k-1 részhalmaz egyedei pedig a tanító adatbázis. Ennek eredményeképpen minden egyed pontosan egyszer volt kiértékelő.

Osztályozási feladat kiértékelési metrikái

-accuracy: helyes/helytelen eset

Ha jobban szeretnénk a teljesítményt mérni akkor érdemes megkülönböztetni:

-igaz pozitív

-hamis negatív

-hamis pozitív

-igaz negatív

Pontosság: ip/(ip+hp) (hány százalékban volt igaza)

Fedés: ip/(ip+hn) (hányat talált meg fedett le a rendszer)

F1 mérték: (2*pontosság*fedés)/(pontosság+fedés)

(pontosság fedés harmonikus közepe)

Regressziós feladat kiértékelési metrikái:

Leggyakrabban használt kiértékelési átlagos négyzetes hiba MSE

MSE: A predikált és a tényleges értékek közötti különbségek négyzeteinek átlagát számolja ki. Ez azt jelenti, hogy minden predikált érték és az annak megfelelő valós érték közötti különbséget négyzetre emeli (így szigorúbban bünteti a nagyobb hibákat), majd ezeket a négyzetes különbségeket átlagolja az egész adathalmazon.

Lineáris regresszió

A lineáris regressziónál egyetlen diszkriminancia függvényt tanulunk és annak kimenete lesz majd a predikciónk. (sztochasztikus gradiens regresszor, regressziós support vector machine SVR)

(A lineáris gépek alapvetően regressziót csinálnak, hiszen a diszkriminancia függvény egy folytonos értéket fog predikálni, amit utána osztályozási döntéssé alakítunk)

Regressziós fák

A modellje egy döntési fa és a leveleiben egy konstans érték vagy egy lineáris regressziós modell van. Az utóbbi esetben minden levélen lineáris regresszós modell van.

Regressziós KNN

A regressziós KNN a k szomszéd tanító adatbázisának célértékének az átlagát fogja predikálni.

5. Túltanulás/Túláltalánosítás

A gyakorlatban a tanító adatbázis és a célváltozó is gyakran zajos. Emiatt nem tudjuk pontosan becsülni.

Túláltalánosítás

A tanító adatbázison mért hibákat nevezzük torzításnak és ha nagy a torzítás az azt jelenti, hogy a gépi tanulási megközelítéssel nem tudtuk elég jól leírni az egyedeket.

Túltanulás

A tanító adatbázison veszünk véletlenszerű részhalmazokat és ezeknek a különbözőségét varienciának hívjuk. Hegy gépi tanulásnál nagy a variencia az azt jelenti, hogy túl jól illeszkedik a zajos tanító egyedre, túltanul.

Metaparamétrek finomhangolása

Az adathalmazt 3 részre osztjuk, tanító, validációs és kiértékelő. A kiértékelő adatbázist csak a végső modell kiértékelésekor használjuk. A metaparaméterek pontossága miatt a tanító adatbázison tanítunk a validációs halmazon pedig kiértékelünk.