IB031 Projekty

Jaroslav Čechák

15. dubna 2019

1 Zadání projektu

Projekty se vypracovávají ve skupinách po max. 3 studentech. Každý projekt ve svém názvu vymezuje skupinu modelů/algoritmů, které se v daném projektu budou používat. Každý student ve skupině provede následující.

1. vybere jeden dataset
2. udělá explorační analýzu dat nad svým datasetem
3. udělá předzpracování svého datasetu
4. vybere jeden konkrétní model/algoritmus pro strojové učení spadající do oblasti vymezené názvem tématu
5. sepíše krátké vysvětlení fungování svého modelu/algoritmu
6. natrénuje svůj vybraný model na všech datasetech vybraných ve skupině
7. provede vyhodnocení všech modelů/algoritmů na svém datasetu
8. sepíše krátké shrnutí výsledků z vyhodnocení

2 Popis jednotlivých úkolů

2.1 Výběr datasetu

Výběr datasetu je čistě na vás. Můžete vybrat veřejně dostupný z internetu nebo klidně i vlastní, na kterém pracujete v rámci laboratoře/semináře/závěrečné práce. Dataset nesmí být triviální (např. Iris) ani příliš „čistý“, aby bylo potřeba předzpracování. Zároveň by měl být dataset rozumně veliký, aby se dal zpracovat na stroji s 8GB operační paměti. Dataset by měl být ideálně součásti odevzdaného projektu, ale v případech, kdy se bude jednat o neveřejná data, jsem ochotný udělat výjimku. Pár tipů, kde hledat datasety:

- https://archive.ics.uci.edu/ml/datasets.php
- https://www.kaggle.com/
- https://www.fi.muni.cz/adaptivelearning/?a=data
- https://guides.library.cmu.edu/machine-learning/datasets
- https://toolbox.google.com/datasetsearch
- https://www.google.cz/

2.2 Explorační analýza

Prozkoumejte dataset, tj. podívejte se, kolik je v datasetu dat a jaká jsou, kolik a jakého typu jsou hodnoty jednotlivých sloupců, jak spolu jednotlivé položky korelují. Výstup této analýzy budou typicky tabulky a grafy. Svá pozorování okomentujte pár větami.

2.3 Předzpracování

Připravte dataset tak, aby se na něm mohly učit jednotlivé modely/algoritmy. Do tohoto kroku patří veškerá manipulace s daty. Může se jednat např. o převody datových typů (např. na factor), práce s chybějícími hodnotami, škálování a normalizace, feature selection, feature extraction, feature engineering, rozdělení na trénovací a testovací množinu, bootstrapping, resampling. Ne všechny vyjmenované věci je potřeba udělat, záleží na datech a modelech/algoritmech. Když už se k nějakému předzpracování rozhodnete, stručně okomentuje co a proč děláte.

2.4 Výběr modelu/algoritmu

Tento krok velice záleží na oblasti zadané názvem tématu. Můžete využít něco ze cvičení, ale ideálně se poohlédněte po něčem novém. Při hledání můžete využít např. následující odkazy.

- `https://topepo.github.io/caret/available-models.html`
- `http://cran.r-project.org/web/views/MachineLearning.html`
- `https://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R`
- `https://www.google.cz/`

2.5 Vysvětlení modelu/algoritmu

Vysvětlení modelu / fungování algoritmu bude v rozsahu cca jednoho odstavce. Cílem je představit techniku a stručně a výstižně popsat její fungování[1].

2.6 Natrénování modelu

Natrénujte svůj vybraný model na všech datasetech ve svojí skupině. Přesný způsob trénování je na vás. Ve většině případů si vystačíte s funkcí `train`, ale některé modely, které nejsou v balíku `caret`, budou vyžadovat jiný způsob trénovaní. Zkuste také vhodně zvolit parametry modelu.

2.7 Vyhodnocení modelu/algoritmu

Na svém vybraném datasetu proveďte porovnání všech naučených modelů ve skupině. Zvolte vhodné metody (míry) pro vyhodnocení modelu v závislosti na řešení úloze (shlukování, klasifikace, regrese, ...). Volbu míry stručně zdůvodněte a okomentujte, co hodnotí.

2.8 Shrnutí výsledků

V pár větách shrňte výsledky z vyhodnocení modelů/algoritmů jak na konkrétních datasetech tak napříč svými datasety ve skupině. Zejména zajímavá jsou zjištění, který model/algoritmus funguje nejlépe na konkrétním datasetu a ideálně zdůvodnění proč. Stejně tak zda jeden model/algoritmus je obecně lepší na všech zkoumaných datasetech.

[1] `https://pbs.twimg.com/media/DTl0aT8VwAAgHHv.jpg`

3 Odevzdání

Hotový projekt odevzdáte jako jediný zip nebo tar.gz archiv do odevzdávárny v ISu (https://is.muni.cz/auth/el/fi/jaro2019/IB031/ode/ode_projekt). Archiv bude obsahovat:

- jediný soubor ve formátu RMakrdown, kde bude veškerý kód (v jazyce R) proložený komentáři a popisnými texty,
- PDF soubor s reportem vygenerovaným z RMarkdownu a
- všechny použité datasety (pokud jsou veřejné).

Deadline pro odevzdání projektu je začátek zkouškového období, tzn. **23. 5. 2019 23:59**.

4 Hodnocení

Níže je rubrika shrnující co a jak budu na projektech hodnotit.

popis požadavku	body
každý student vybral netriviální dataset vhodně k řešené úloze	2
projekt obsahuje explorační analýzu pro všechny datasety	2
vhodné předzpracování dat podle typu řešené úlohy a vybraných datasetů	2
propracované předzpracování dat s využitím pokročilých technik	2
každý student vybral model/algoritmus z rodiny určené tématem projektu	2
projekt obsahuje krátký popis pro každý vybraný model/algoritmus	2
všechny modely natrénované na všech vybraných datasetech	1
vhodná volba parametrů při trénování a jejich kontrola pomocí validace	4
vyhodnocení modelů/algoritmů pro každý dataset pomocí vhodně zvolené míry	4
projekt obsahuje krátké shrnutí výsledků na jednotlivých datasetech i napříč datasety	2
projekt obsahuje vysvětlující komentáře dokumentující jednotlivá rozhodnutí v projektu	2
správná metodologie učení a vyhodnocování modelů/algoritmů	5

Za projekt můžete získat i bonusové body. Rubrika níže uvádí za co a kolik bonusových bodů můžete získat v projektu získat.

popis požadavku	body
porovnání modelů s nějakým naivním „baseline“ modelem	2
za každý vybraný model/algoritmus, který se neprobíral na cvičení	2
využití techniky feature engineering/extraction/selection v předzpracování	4
porovnání několika různých mír k měření kvality modelu	4