

Pokyny k vypracovaniu projektov

2-INF-150: Strojové učenie, Zima 2015

Vyučujúci: Vladimír Boža

Spoločné pokyny pre všetky varianty projektov

Základnou úlohou v rámci projektu je:

- vybrať si nejaký problém, na ktorý sa dajú aplikovať metódy strojového učenia (môže byť takmer čokoľvek, obzvlášť vhodné je niečo, čo potrebujete urobiť do diplomovej práce),
- naimplementovať jednu alebo viacero variant riešenia problému,
- otestovať implementáciu na skutočných alebo simulovaných dátach.

Prednášajúci je dudroš a nemá rád projekty, ktoré sú nezaujímavé a pôsobia štýlom “aby niečo bolo”, resp. “zobral som tento štandardný nezaujímavý dataset z UCL a použil metódy, ktoré sme sa učili a dostal som takéto výsledky (ako stovka ďalších iných ľudí predomnou)”. Zaujímavý projekt buď pracuje so zaujímavými dátami, alebo skúša cutting edge metódy. **Zopár nápadov na projekty sa nachádza na konci týchto pokynov.**

Je odporúčané skonzultovať vhodnosť projektu s prednášajúcim do konca semestra.

Termíny: Samotný projekt treba odovzdať najneskôr **3 dni pred skúškou**, maximálne do **29. 1. 2016**. Zrojové kódy (ideálne link na github repozitár, s dobrým README) a prípadný písomný zhrnutie odovzdávajú emailom na adresu prednášajúceho.

Orientačné hodnotenie:

- Vhodnosť výberu problému: 20%
- Výber a správne použitie metód: 30%
- Prevedenie projektu (implementácia, kvalita reportu, ...): 50%

Opisovanie. Tak ako vo zvyšku tohto kurzu, opisovanie (či už od spolužiakov alebo z internetu) bude prísne postihované – vid' formálne upozornenie na stránkach predmetu. Keďže ide o veľkú časť hodnotenia, v prípade zistenia opisovania nemožno rátať s úspešným ukončením predmetu. **Za opisovanie pokladáme aj ak odovzdáte prácu, ktorú ste už v podobnej forme odovzdali na iný predmet.**

Variant A – zhrnutie v štýle konferenčného článku

Písomné zhrnutie má byť v rozsahu 5-15 strán (nehodnotí sa dĺžka, ale obsah) a je hlavným podkladom pre hodnotenie projektu. Report musí obsahovať najmenej:

- Nadpis a meno autora projektu.
- Abstrakt (cca 1-2 odstavce) sumarizujúci cieľ, použité metódy a výsledky projektu.
- Úvod do problematiky: popis úlohy, odkiaľ ste získali dáta a ich stručná charakteristika (napr. veľkosť trénovacej a testovacej množiny, koľko pozitívnych a negatívnych príkladov, priemernú veľkosť každej inštancie, popis atribútov, priemery a štandardné odchýlky najzaujímavejších

- atribútov a pod.).
- Popis a zdôvodnenie použitých metód.
- Stručný popis technických detailov implementácie (v akom jazyku, aké knižnice ste použili a prečo, s akými technickými problémami ste sa stretli pri riešení).
- Experimentálne vyhodnotenie: spôsob vyhodnocovania úspešnosti, porovnanie s niečím iným, tabuľky, grafy, ich interpretácia a závery.
- Záver: Myslíte si, že bol projekt úspešný alebo neúspešný? Prečo? Keby ste projekt robili znovu, čo by ste urobili ináč? Aké problémy by sa dali riešiť podobným spôsobom?

Variant B – webová aplikácia

Tento variant je vhodný vtedy pokiaľ vami použité techniky dávajú zaujímavé výsledky a dajú sa používať interaktívne.

Webová aplikácia musí bežať u vás a navyše musí obsahovať podstránku so zhrnutím projektu v rozsahu asi 1-2 A4. Toto zhrnutie by malo obsahovať:

- Zhrnutie cieľu projektu.
- Link na dáta ktoré ste použili.
- Popis použitých metód.
- Stručné vyhodnotenie úspešnosti (ak sa dá urobiť).

Variant C – Kaggle súťaž

Pokiaľ sa vám podarí v niektorej z Kaggle súťaží (<https://www.kaggle.com/competitions>), ktoré majú finančnú odmenu, dostať na aspoň 10 miesto (stačí byť na 10 mieste aspoň chvíľu) a od začiatku súťaže ubehli aspoň 2 týždne, tak za projekt môžete získať $52 - 2 * X$ bodov (kde X je vaša pozícia vo výsledkovke). Podmienkou získania bodov je zaslanie stručného zhrnutia použitých metód.

Samozrejme, pokiaľ sa rozhodnete robiť Kaggle súťaž a neskončíte vysoko, stále môžete odovzdať dlhší report.

Zopár nápadov na projekty:

- z obrázku určiť otočenie fotografie (na šírku/na výšku)
- predikovať výsledok tímovej online hry (napr. Dota 2) z vhodne pripravených dát (napr. údaje o hráčoch na oboch stranách, ...)
- rôzne predikcie v štátnych open dátoch (<http://datanest.fair-play.sk/datasets>), resp. otvorených zmluvách (<http://www.otvorenezmluvy.sk/data/>), napr. zaujímavé by bolo automaticky vyhľadávať predražené nákupy bežného tovaru
- rôzne problémy okolo spracovania prirodzeného jazyka (vhodné datasety napríklad tu <http://nlp.stanford.edu/research.shtml>)