

# IB031 Úvod do strojového učení

## Pokyny k projektům

8. března 2021

### 1 Zadání projektu

Projekty se vypracovávají ve skupinách po max. 3 studentech a všichni studenti musí být ze stejné seminární skupiny. Každý projekt vymezuje dataset a případné další podmínky na vypracování, ve zbytku rozhodnutí máte úplnou volnost. Typicky každý student ve skupině provede následující.

1. zapojí se do vypracování explorační analýzy dat
2. zapojí se do předzpracování dat
3. vybere si jeden konkrétní model pro strojové učení
4. sepíše krátké vysvětlení fungování svého modelu
5. natrénuje svůj vybraný model na předzpracovaném datasetu
6. vyhodnotí svůj model na datasetu
7. zapojí se do sepisování krátkého shrnutí výsledků z porovnání

### 2 Popis jednotlivých úkolů

#### 2.1 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. Ke každému datasetu je vždy i dokument (typicky vědecký článek), který dataset popisuje. Ten je dobré přečíst před započítím analýzy. Výstup této analýzy budou typicky tabulky a grafy. Svá pozorování okomentujte pár větami.

#### 2.2 Předzpracování

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

## 2.3 Výběr modelu

Výběr modelu je omezený pouze typem úlohy (regrese, klasifikace, shlukování, detekce anomálií) vašeho konkrétního zadání, jinak máte volný výběr. Můžete zvolit model z knihovny `scikit-learn`, ale klidně můžete použít jinou knihovnu nebo si dokonce implementovat vlastní model.

## 2.4 Vysvětlení modelu

Vysvětlení modelu bude v rozsahu cca jednoho odstavce. Cílem je představit techniku a stručně a výstižně popsat její fungování<sup>1</sup>.

## 2.5 Natrénování modelu

Natrénujte svůj vybraný model na předzpracovaném datasetu. Přesný způsob trénování je na vás, ale měl by být porovnatelný napříč modely. Tedy není dobré, aby se každý model trénoval na jiné podmnožině dat. Model se nemusí učit na všech sloupcích, může si vybrat jen jejich podmnožinu. Součástí trénování bude i ladění hyperparametrů.

## 2.6 Vyhodnocení modelu

Na svém datasetu proveďte vyhodnocení modelu pomocí vhodně zvolené míry v závislosti na řešené úloze (regrese, klasifikace, shlukování, detekce anomálií). Volbu míry stručně zdůvodněte a okomentujte, co měří.

## 2.7 Shrnutí výsledků

V pár větách srovnajte výsledky jednotlivých modelů a shrňte výsledky z vyhodnocení. Zejména zajímavá jsou zjištění, který model funguje nejlépe a ideálně zdůvodnění proč. Stejně tak zda je některý z modelů robustnější při volbě různých rozdělení dat na trénovací a testovací množinu.

# 3 Odevzdání

Své výsledky budete dvakrát prezentovat na cvičení a zároveň odevzdáte finální verzi do odevzdávacího v ISu.

První prezentace proběhne na 9. cvičení (týden 26. 4. až 2. 5.). Během této prezentace krátce představíte dataset, výsledky explorační analýzy a kam jste se s vypracováním projektu dostali. Zároveň byste měli mít rozmyslené, jak chcete dataset předzpracovat a jaké modely budete učit. Prezentace bude opravdu krátká, pouze několik minut.

Druhá prezentace proběhne na posledním 13. cvičení (týden 24. 5. až 30. 5.). Na této prezentaci už představíte kompletní řešení svého projektu, vaše výsledky a pozorování. Zároveň každý z členů týmu řekne jednu věc, kterou si z projektu odnesl. Prezentace bude maximálně na 10 minut.

Hotový projekt odevzdáte jako jediný zip nebo tar.gz archiv do odevzdávacího v ISu. Archiv bude obsahovat:

- jediný IPython notebook, kde bude veškerý kód proložený komentáři a popisnými texty,
- PDF dokument vygenerovaný z vyhodnoceného notebooku,
- případné další zdroje dat, které jsou používali.

Deadline pro odevzdání projektu je **konec 13. cvičení**.

<sup>1</sup><https://pbs.twimg.com/media/DT10aT8VwAAGHHv.jpg>

## 4 Hodnocení

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

popis požadavku	body
projekt obsahuje explorační analýzu datasetu	2
vhodné předzpracování dat podle typu řešené úlohy a vybraných modelů	2
propracované předzpracování dat s využitím pokročilých technik	4
projekt obsahuje krátký popis pro každý vybraný model	2
všechny modely natrénované na datasetu	1
vhodná volba parametrů modelů a jejich ladění	4
vyhodnocení modelů pomocí několika vhodně zvolených mír	4
porovnání modelů s nějakým naivním „baseline“ modelem	2
projekt obsahuje krátké shrnutí výsledků	2
projekt obsahuje vysvětlující komentáře dokumentující jednotlivá rozhodnutí v projektu	2
správná metodologie učení a vyhodnocování modelů	5