Posudek oponenta diplomové práce Bc. Jiřího Chmele

Ve své práci nazvané **Machine learning for prediction of energy in condensed matter physics** se student zabývá aplikacemi metod strojového učení na problémy z oblasti elektronové struktury pevných látek.

V první ze čtyř kapitol jsou popsány vybrané metody strojového učení (regresní metody a metody využívající umělé neuronové sítě) a postupy používané pro posouzení kvality modelů, které jsou těmito metodami zkonstruovány. V druhé kapitole jsou diskutovány specifické aspekty aplikací strojového učení na problémy týkající se elektronové struktury pevných látek a z ní odvozených fyzikálních vlastností. Především jde o výběr a organizaci charakteristik pevných látek, takzvaných deskriptorů, které slouží jako vstupní data pro metody strojového učení. Cílem aplikací (numerických experimentů) popsaných v následujících dvou kapitolách je, velmi zjednodušeně řečeno, určení totální energie krystalu nějaké sloučeniny na základě znalosti totálních energií mnoha jiných v nějakém smyslu příbuzných sloučenin. Deskriptor pro takovou aplikaci potom zahrnuje relevantní charakteristiky těchto sloučenin (strukturní data) a/nebo jejich konstituentů (atomová čísla, elektronegativity, . . .), na kterých totální energie závisí.

Práce je napsána srozumitelně, nicméně se zdá, že tu a tam nějaká informace vypadla. Například na straně 29 je zmíněn *generalization error*, ale nikde není popsáno, co přesně tento pojem znamená. Zároveň mi místy chybí popis motivace pro některé kroky (například proč byla zvolena tato metoda a ne jiná) – k tomu se vrátím podrobněji v otázkách. Ze samotného textu také není vždy jasné, co je výsledek tvůrčí práce studenta, co je opakování již dříve publikovaných numerických experimentů a kde už začíná jejich rozšíření. Je nicméně zřejmé, že student zadané problematice dobře porozuměl, práci věnoval hodně času a při rozsáhlých numerických experimentech postupoval pečlivě a systematicky. Osobně oceňuji mimo jiné podrobnou analýzu provedenou v kapitole 3 (3.1.4 a 3.1.5). Podle mého názoru tedy bylo **zadání práce beze zbytku splněno**. Diplomovou práci **doporučuji k obhajobě** a navrhuji ji hodnotit známkou **B – velmi dobře**. Jak jsem popsal výše, k hodnocení výborně, myslím, trochu chybí.

Pro diskusi během obhajoby mám na studenta následujících několik otázek:

1. Na straně 44 se píše *It comes from quantum mechanics that the physical features are correlated in terms of the Pearson correlation coefficient of two 82-dimensional feature vectors*. To je poměrně obecné tvrzení. Bylo by možné rozebrat trochu podrobněji, odkud zmíněná korelace pochází?

Fyzikální ústav Akademie věd České republiky, v. v. i.
Na Slovance 1999/2 182 21 Praha 8 Česká republika
Tel.: +420 266 052 110
E-mail: info@fzu.cz
Web: www.fzu.cz

2. Pro optimalizaci *hyperparametrů* byl použit grid search, přičemž je zmíněno, že je to výpočetně velmi náročné (1.3.3., 4.2.1.3). Nebylo by možné/vhodné použít nějakou pokročilejší metodu optimalizace?

3. Deskriptor *ngram* byl v kapitole 4 použit ve spojení s *kernel ridge regresion*, zatímco *SOAP* deskriptor byl použit ve spojení s *neuronovými sítěmi*. Jaký byl důvod kombinovat s různými deskriptory i různé metody? Pro porovnání kvality deskriptorů bych považoval vhodnější použít oba se stejnou metodou pro minimalizaci rozdílů mimo definice deskriptorů.

V Praze 19. května 2022

Mgr. Jindřich Kolorenč, PhD.
kolorenc@fzu.cz
+420 2 6605 2914

Fyzikální ústav
Akademie věd České
republiky, v. v. i.

Na Slovance 1999/2
182 21 Praha 8
Česká republika

Tel.: +420 266 052 110
E-mail: info@fzu.cz
Web: www.fzu.cz