Projekt

Cześć III (Końcowa)

Grupy tydzień nieparzysty (nr 1, 2, 4, 6):
- Wprowadzenie: ~~09 maja 2025~~ 25 kwietnia 2025
- Termin wysłania projektu końcowego: 25 maja 2025
- Termin oddania: 26-28 maja 2025

Grypy tydzień parzysty (nr 3, 5, 7):
- Wprowadzenie: 23 maja 2025
- Termin wysłania projektu końcowego: 6 czerwca 2025
- Termin oddania: 9 czerwca 2025

Celem trzeciej (ostatniej) części projektu jest stworzenie i zaprezentowanie kompletnego projektu, na który będą składać się wszystkie wcześniejsze cześci oraz nowe elementy związanie z optymalizacją modeli ML.

Część implementacyjna

W ramach prac na elematami optymalizacji skup się na następujących zadaniach:

1. Cross-validation i ewaluacja modelu

- Przeprowadź 3-krotną walidację krzyżową (zobacz `KFold` lub `StratifiedKFold` w przypadku klasyfikacji).
- Sprawdź czy wyniki są podobne na wszystkich podzbiorach. Jeśli nie, co to oznacza?
- *Zintegruj metodę ze swoją implementacja regresji.
- Patrz: https://scikit-learn.org/stable/modules/cross_validation.html

2. Wykresy zbieżności i analiza błędów

- Dla modelu regresji liniowej lub logistycznej zweryfikuj, czy istnieje problem nadmiernego lub niedostatecznego dopasowania. Zwiększ złożoność modelu poprzez dodanie dodatkowych cech (np. PolynomialFeatures) i ponownie sprawdź, czy problem zachodzi.
- *Dla własnej implementacji regresji (z gradient descent), stwórz wykres funkcji kosztu na podzbiorach treningowym oraz testowym względem epoki. Zbadaj, czy następuje zbieżność (spadek błędu) w czasie uczenia modelu.
- Przeanalizuj, czy występuje overfitting lub underfitting na podstawie wykresu.
- Sprzwdź również jaki wpływ ma dodanie cech oraz ograniczenie cech (np. wybór tylko cześci kolum oraz zmniejszenie liczebności zbioru)

- Patrz: https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

3. Regularyzacja L1 i L2

- Dodaj regularyzację do modeli regresji liniowej lub logistycznej:
- Porównaj dwie metody `Ridge` (L2), `Lasso` (L1)
- Porównaj wyniki z modelami bez regularyzacji.
- Wyjaśnij wpływ regularyzacji na wagę cech i potencjalne przetrenowanie.
    - Sprawdź wartości wag przed i po regularyzacji.
- *Zaimplementuj metodę regularyzacji dla własnej implementacji regresji (podobnie jak wcześniej użyj możesz wprost dodać obliczoną pochodną do formuły updatującej wagi).
- Patrz: https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization

4. Usprawnienie danych – balansowanie zbiorów

- Zastosuj metody radzenia sobie z niezbalansowanymi danymi:
    - **Oversampling** (np. `imblearn.over_sampling.SMOTE`)
    - **Undersampling**
- Porównaj modele trenowane na oryginalnym i zbalansowanym zbiorze.
- Zastosuj metryki ewaluacji adekwatne dla problemu niezbalansowanego, np. **precision**, **recall**, **f1-score**.
- Uwaga. Jeśli twój model rozwiązuje problem regresji, a nie klasyfikacji, możesz wykonać podpunkt na innej kolumnie jak “target”.
- *Zintegruj metodę ze swoją implementacja regresji.
- Patrz: https://datasciencehorizons.com/handling-imbalanced-datasets-in-scikit-learn-techniques-and-best-practices/

5. Optymalizacja hiperparametrów

- Wybierz dwa modele i przeprowadź strojenie hiperparametrów przy użyciu `GridSearchCV` lub innej podobnej techniki
- Wybierz adekawatne parametry dla modelu, np. głębię drzewa (`max_depth`), liczbę sąsiadów (`n_neighbors`), itp.
- Dlaczego przeszukwanie parametrów jest ogólnie trudnym problemem?
- Zaprezentuj najlepsze parametry i ich wpływ na końcowy wynik.
- Patrz: https://scikit-learn.org/stable/modules/grid_search.html

6. Ensemble methods

- Zastosuj proste metody **ensembe**: `VotingClassifier`, `StackingClassifier` **(**lub odpowiedniki dla regresji)
- Postaraj się wybrać takie modele, które we wcześniejszych eksperymentach popełniały błędy na różnych rekordach
- *(na 5.5)* Zaprojektuj oraz zaimplementuj metodę Mixture of Experts
    - Dlaczego jest to rozwiązanie wykorzystywanie w przypadku dużch modeli?
- Patrz: https://scikit-learn.org/stable/modules/ensemble.html#voting-classifie
- Patrz 2: https://scikit-learn.org/stable/modules/ensemble.html#stacked-generalization

* wymaga wykonania Cześć II przynajmniej na ocenę 4.0.

Punktacja:

- Na ocenę 3.0 konieczne jest wykonanie 1 zadań z 6.

- Na ocenę 3.5 konieczne jest wykonanie 2 zadań z 6.

- Na ocenę 4.0 konieczne jest wykonanie 3 zadań z 6.

- Na ocenę 4.5 konieczne jest wykonanie 4 zadań z 6.

- Na ocenę 5.0 konieczne jest wykonanie 5 zadań z 6.

- Na ocenę 5.5 konieczne jest wykonanie 6 zadań z 6 (w tym podpunktu na 5.5).

Część raportowa

Stwórz spójny raport końcowy, który będzie zawierał analizę zbioru danych z Części I, wyniki treningu i ewaluacji modeli z Części II oraz wyniki optymalizacji z Części III.

W ramach prezentacji Cześci III:

1. Przygotuj studim ablacyjne (ablation study), w którym zaprezentujesz, w postaci tabeli, wyniki z uwzględniem metod optymalizacji. O ile to możliwe metody te mają być dodwane akumlacyjnie do modelu. Przykładowo, w tabeli mają być widoczne wyniki dla modelu z regularyzacja, modelu z regularyzacją + metodą emsemble.
2. Opisz model, który dał najlepszy wynik. Postaraj się sprawdzić, czemu akurat ten model okazał się najlepszy.
3. Zaprezentuj wyniki cząstkowe:
    a. zaprezentuj w postaci tabeli uzyskane wyniki walidacji krzyżowej (jeśli wybrałeś zadanie nr 1) oraz dopisz swoje wnioski,
    b. zaprezentuj wykresy zbieżności (jeśli wybrałeś zadanie nr 2), przeanalizuj wuniki oraz dopisz swoje wnioski,
    c. zaprezentuj w postaci tabeli uzyskane wyniki regularyzacji (jeśli wybrałeś zadanie nr 3), wypisz wagi cech oraz dopisz swoje wnioski,

d. zaprezentuj w postaci tabeli uzyskane wyniki przed i po zbalansowaniu danych (jeśli wybrałeś zadanie nr 4), pokaż wyniki metryk oraz dopisz swoje wnioski,
e. zaprezentuj w postaci tabeli uzyskane wyniki przeszukiwania hiper parametrów (jeśli wybrałeś zadanie nr 5) oraz dopisz swoje wnioski,
f. zaprezentuj w postaci tabeli uzyskane wyniki metody ensemble oraz odzielnie wykorzystawnych modeli (jeśli wybrałeś zadanie nr 6) oraz dopisz swoje wnioski.

Uwaga. W przypadku, gdy raport końcowy będzie spójny, kompletny oraz rzetelny, ocena końcowa z kursu zostanie zaproponowana oraz (po akceptacji studenta) wystawiona bez konieczności osobistej prezentacji raportu na zajęciach.