Laboratorium: Metody zespołowe

April 7, 2022

1 Cel/Zakres

- Metody zespołowe:
  - równoległe,
  - sekwencyjne.
- Hard/soft voting.
- Bagging.
- Boosting.

2 Przygotowanie danych

```
from sklearn import datasets
data_breast_cancer = datasets.load_breast_cancer(as_frame=True)
```

3 Ćwiczenie

Uwaga: stosuj domyślne wartości parametrów dla użytych klas, chyba, że z opisu danego ćwiczenia wynika inaczej.

1. Podziel zbiór `data_breast_cancer` na uczący i testujący w proporcjach 80:20.
2. Zbuduj ensemble używając klasyfikatorów binarnych, których używałeś(aś) w poprzednich ćwiczeniach, tj.: drzewa decyzyjne, regresja logistyczna, k najbliższych sąsiadów, do klasyfikacji w oparciu o cechy: `mean texture`, `mean symmetry`. Użyj domyślnych parametrów.
3. Porównaj dokładność (accuracy) ww. klasyfikatorów z zespołem z głosowaniem typu *hard* oraz *soft*.
4. Zapisz rezultaty jako listę par (*dokładność_dla_zb_uczącego*, *dokładność_dla_zb_testującego*) dla każdego z w/w klasyfikatorów (razem 5 elementów) i umieść ją w pliku Pickle o nazwie `acc_vote.pkl`

   5 pkt

   Zapisz klasyfikatory jako listę w pliku Pickle o nazwie `vote.pkl` (5 obiektów).

   2 pkt
5. Wykonaj na zbiorze uczącym wykorzystując 30 drzew decyzyjnych:
   - Bagging,

- Bagging z wykorzystaniem 50% instancji,
- Pasting,
- Pasting z wykorzystaniem 50% instancji, oraz
- Random Forest,
- AdaBoost,
- Gradient Boosting.

Dlaczego Random Forest daje inne rezultaty niż Bagging + drzewa decyzyjne?

6. Oblicz dokładności oraz zapisz je jako listę par (*dokładność_dla_zb_uczącego*, *dokładność_dla_zb_testującego*) dla każdego z ww. estymatorów (razem 7 elementów) w pliku Pickle o nazwie `acc_bag.pkl`.

   7 pkt

   Zapisz klasyfikatory jako listę w pliku Pickle o nazwie `bag.pkl`

   2 pkt

7. Przeprowadź sampling 2 cech z wszystkich dostepnych bez powtórzeń z wykorzystaniem 30 drzew decyzyjnych, wybierz połowę instancji dla każdego z drzew z powtórzeniami.

8. Zapisz dokładności ww estymatora listę : `dokładność_dla_zb_uczącego`, `dokładność_dla_zb_testującego` w pliku Pickle `acc_fea.pkl`.

   2 pkt

   Zapisz klasyfikator jako jednoelementową listę w pliku Pickle o nazwie `fea.pkl`

   1 pkt

9. Sprawdź, które cechy dają najwięszą dokładność. Dostęp do poszczególnych estymatorów, aby obliczyć dokładność, możesz uzyskać za pmocą: `BaggingClasifier.estimators_`, cechy wybrane przez sampling dla każdego z estymatorów znajdziesz w: `BaggingClassifier.estimators_features_`. Zbuduj ranking estymatorów jako `DataFrame`, który będzie mieć w kolejnych kolumnach: dokładność dla zb. uczącego, dokładnośc dla zb. testującego, lista nazw cech. Każdy wiersz to informacje o jednym estymatorze. `DataFrame` posortuj malejąco po wartościach dokładności dla zbioru testującego i uczącego oraz zapisz w pliku Pickle o nazwie `acc_fea_rank.pkl`

   5 pkt

4 Prześlij raport

Prześlij plik o nazwie `lab6.py` realizujący ww. ćwiczenia.

Sprawdzane będzie, czy skrypt Pythona tworzy wszystkie wymagane pliki oraz czy ich zawartość jest poprawna.