Automatyczne Uczenie Maszynowe
Projekt 1

Wiktoria Boguszewska i Mateusz Zacharecki

20 listopada 2024

Problemy

Problemy do rozważenia:

1. ile iteracji każdej metody potrzebujemy żeby uzyskać stabilne wyniki optymalizacji,
2. określenie zakresów hiperparametrów dla poszczególnych modeli,
3. tunowalność poszczególnych algorytmów,
4. czy technika losowania punktów wpływa na różnice we wnioskach w punkcie 3. dotyczącym tunowalności algorytmów, czy występuje bias sampling.

Wprowadzenie

Cel

Celem projektu jest przeanalizowanie tunowalności hiperparametrów 3 wybranych algorytmów uczenia maszyno- wego na co najmniej 4 zbiorach danych. Do tunowania modeli należy wykorzystać min. 2 różne techniki losowania punktów.

Dane

W projekcie wykorzystujemy 4 różne zbiory danych dla problemu klasyfikacji binarnej:

- Breast Cancer,
- Credit,
- Blood Transfusion,
- Bank marketing.

Metody sampligu

Metody samplingu

Główna część projektu polegała na zastosowaniu różnych metod tunowalności hiperparametrów. Wykonaliśmy to dla 3 algorytmów: regresja logistyczna, KNN oraz XGBoost. Zastosowane przez nas metody samplingu to:

- Random Search,
- Bayes Search,
- Grid Search.

Wyniki

Wyniki dla poszczególnych metod

MODEL	LOGISTYCZNA				KNN				XGBOOST			
ZBIÓR	cancer	credit	blood	bank	cancer	credit	blood	bank	cancer	credit	blood	bank
TRAIN ACCURACY	0.9736	0.7175	0.7843	0.8827	0.967	0.7138	0.8027	0.8866	0.9363	0.7025	0.7792	0.8827
TEST ACCURACY	0.9825	0.68	0.72	0.8928	0.9912	0.705	0.7533	0.8961	0.9561	0.685	0.74	0.8928

Tabela: Wyniki otrzymane za pomocą metody Random Search.

MODEL	LOGISTYCZNA				KNN				XGBOOST			
ZBIÓR	cancer	credit	blood	bank	cancer	credit	blood	bank	cancer	credit	blood	bank
TRAIN ACCURACY	0.9736	0.7213	0.786	0.8836	0.967	0.7175	0.8094	0.8866	0.9758	0.7188	0.796	0.8858
TEST ACCURACY	0.9825	0.685	0.72	0.8961	0.9737	0.685	0.7533	0.8961	0.9649	0.685	0.76	0.9028

Tabela: Wyniki otrzymane za pomocą metody Bayes Search.

MODEL	LOGISTYCZNA				KNN				XGBOOST			
ZBIÓR	cancer	credit	blood	bank	cancer	credit	blood	bank	cancer	credit	blood	bank
TRAIN ACCURACY	0.9692	0.72	0.786	0.8836	0.967	0.7113	0.8044	0.8852	0.967	0.7075	0.7994	0.8852
TEST ACCURACY	0.9912	0.69	0.72	0.895	0.9737	0.705	0.7467	0.8983	0.9649	0.72	0.7533	0.8961

Tabela: Wyniki otrzymane za pomocą metody Grid Search.

Analiza wyników

Liczba iteracji

- 50 iteracji każdej z metod.
- 30 iteracji wewnątrz metody bayesian search.
- Najlepsza metoda: **bayesian search**.

Zakres hiperparametrów

Algorytm	Hiperparametr	Typ	Dolna granica	Górna granica	Rozkład
LogisticRegression	**C**	numeric	10^{-10}	10^{10}	loguniform
	l1_ratio	numeric	0	1	uniform
KNeighborsClassifier	**n_neighbors**	integer	1	30	uniform
	weights	discrete	–	–	uniform
	p	integer	1	2	uniform
XGBClassifier	booster	discrete	–	–	uniform
	eta	numeric	10^{-10}	1	loguniform
	n_estimators	integer	1	500	uniform
	max_depth	integer	1	10	uniform
	subsample	numeric	0.1	1	uniform
	colsample_bytree	numeric	0	1	uniform
	colsample_bylevel	numeric	0	1	uniform
	lambda	numeric	10^{-5}	10^{2}	loguniform
	alpha	numeric	10^{-5}	10^{2}	loguniform
	min_child_weight	numeric	1	10^{7}	loguniform

Tabela: Zakresy hiperparametrów dla metody Random Search oraz Bayes Search.

Tunowalność poszczególnych algorytmów

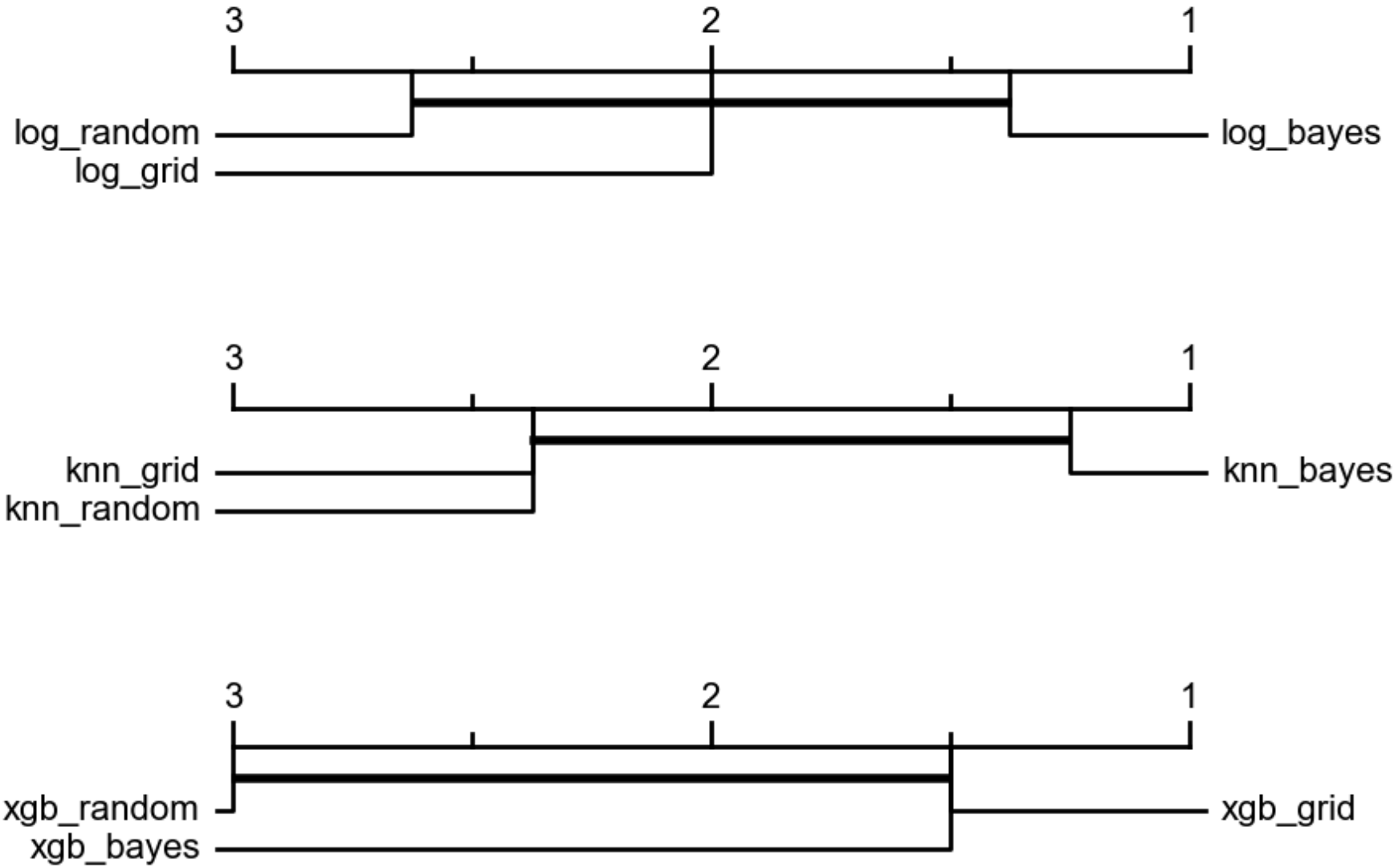
Model	Defaultowe hiperparametry	Tunowalność
Regresja logistyczna	C: 6.563234996980466e-06, l1_ratio: 0.3434560140483249 3	$-3.5014 \cdot 10^{-6}$
kNN	weights: uniform, p: 2, n_neighbors: 18	-0.0004
XGBoost	alpha: 0.0060666419159810 54, booster: gbtree, colsample_bylevel: 0.141194832453028, colsample_bytree: 0.5588264346475861, eta: 2.47082902598744 02e-08, lambda: 2.051302466178794, max_depth: 8, min_child_weight: 5.462765231003678 5, n_estimators: 237, subsample: 0.2477228436396202	0

Tabela: Defaultowe hiperparametry i tunowalność dla metody Random Search.

Testy statystyczne

- Test Manna-Whitneya do porównania różnic pomiędzy technikami losowania hiperparametrów.
- Porównanie metod randomized search oraz bayesian search.
- Różnica wyników z obu modeli jest istotna statystycznie, a wyniki pochodzą z różnych rozkładów.
- Występuje bias sampling.

Critical Difference Diagrams



Bibliografia

[1] Philipp Probst, Anne-Laure Boulesteix, Bernd Bischl; Tunability: Importance of Hyperparameters of Machine Learning Algorithms

Dziękujemy za uwagę