Badanie jadalności grzybów

06-DUMAUI0 2022/SZ

Cel projektu

Celem projektu było stworzenie modelu, który przewidzi na podstawie cech czy grzyb jest jadalny czy trujący.

Dane

Dane pochodzą z wyzwania „Mushroom classification challenge” na platformie Gonito.pl (link: https://gonito.net/challenge/mushrooms).

Po pobraniu danych uzyskano 6465 przykładów, które podzielono na zbiór uczący (4848 przykłady) i zbiór testowy (1617 przykładów). Dodatkowo do dokładnej ewaluacji modelu użyto dodatkowego zestawu danych z powyższego repozytorium (792 przykładów).

Dane zawierały same zmienne kategorialne, dlatego, żeby umożliwić stworzenie z nich modelu, dane te zostały przedstawione w kodowaniu „jeden z N”.

Modele

W projekcie porównano działanie 5 modeli:

- Regresja logistyczna wielomianowa 2. stopnia. Zastosowano domyślną regularyzację L2, parametrem regularyzacji 10.
- Naiwny klasyfikator bayesowski Gaussian. Zostały użyte parametry domyślne.
- Klasyfikacja wektorowa z jądrem rbf, z parametrem regularyzacji 10 i parametrem gamma 0.1.
- Klasyfikator K-najbliższych sąsiadów z parametrem k-najbliższych sąsiadów równym 3.
- Klasyfikator sieci neuronowej z funkcją aktywującą relu, z ilością węzłów 10 i funkcją do optymalizacji wagi „lbfgs”.

Ewaluacja

Do ewaluacji wykorzystano metryki *accuracy, precision, recall* i *F1-score*. Wyniki ewaluacji przedstawia poniższa tabelka:

Model	Accuracy	Precision	Recall	F1-score
Naiwny klasyfikator bayesowski	0.91	0.92	0.91	0.91
Regresja Logiczna wielomianowa	1.00	1.00	1.00	1.00
Klasyfikacja	1.00	1.00	1.00	1.00

wektorowa				
Klasyfikator K-najbliższych sąsiadów	1.00	1.00	1.00	1.00
Klasyfikator sieci neuronowej	1.00	1.00	1.00	1.00

Wnioski

JJedynie naiwny klasyfikator bayesowski nie poradził sobie z zadaniem, uzyskując ze wszystkich metryk wynik poniżej 1.0. Wynika to najpewniej z niedopasowania modelu. Pozostałe modele poradziły sobie z tym zadaniem bardzo dobrze. Jedynych znaczących różnic trzeba się doszukiwać w czasie tworzenia modelu gdzie „Regresja logiczna wielomianowa” i „Klasyfikacja wektorowa” tworzyły się najwolniej a „Naiwny klasyfikator bayesowski” i „Sieci neuronowe” stworzyły model najszybciej.