ЗАСТОСУВАННЯ МЕТРИЧНОГО АЛГОРИТМУ ДЛЯ АВТОМАТИЧНОЇ КЛАСИФІКАЦІЇ ОБ'ЄКТІВ МЕДИЧНОЇ ДІАГНОСТИКИ

Шевченко Сергій Сергійович, студент групи 365*
Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ»

Мета класифікації складається у прогнозуванні для об'єкту мітки класу, яка являє собою вибір з певного списку можливих варіантів. Метод k-найближчих сусідів – метричний алгоритм для автоматичної класифікації об'єктів або регресії. У разі використання методу для класифікації об'єкт присвоюється тому класу, який є найбільш поширеним серед k-сусідів даного елементу, класи яких вже відомі.

Алгоритм k найближчих сусідів є досить прямолінійним і може бути описаним наступними кроками:

– Обрати число k і метрику відстані(найчастіше використовується Евклідова відстань).

– Знайти k найближчих сусідів зразка, який ми хочемо класифікувати.

– Присвоїти мітку класу, до якого відноситься більша частина сусідів зразка.

Залежно від обраної метрики відстані, алгоритм знаходить в тренувальному наборі даних k зразків, які є найбільш схожими до об'єкту, клас якого треба спрогнозувати. Мітка класу нової точки даних потім визначається класом, до якого відноситься більша кількість його сусідів.

У роботі використано базу даних Pima Indians Diabetes DataBase (рис. 1). Кожен приклад представляє окремих пацієнтів та їхні різні медичні характеристики разом із класифікацією діабету. Кількість екземплярів: 768. Кількість атрибутів – 9 (такі як Pregnancies, PG Concentration, Diastolic BP, Tri Fold Thick, Serum Ins, Body Mass Index, Diabetes Pedigree Function, Age, Diabetes).

	A	B	C	D	E	F	G	H	I
1	Pregnancies	PG Concentration	Diastolic BP	Tri Fold Thick	Serum Ins	BMI	DP Function	Age	Diabetes
2	6	148	72	35	0	33,6	0,627	50	Sick
3	1	85	66	29	0	26,6	0,351	31	Healthy
4	8	183	64	0	0	23,3	0,672	32	Sick
5	1	89	66	23	94	28,1	0,167	21	Healthy
6	0	137	40	35	168	43,1	2,288	33	Sick
7	5	116	74	0	0	25,6	0,201	30	Healthy
8	3	78	50	32	88	31	0,248	26	Sick
9	10	115	0	0	0	35,3	0,134	29	Healthy
10	2	197	70	45	543	30,5	0,158	53	Sick
11	8	125	96	0	0	0	0,232	54	Sick
12	4	110	92	0	0	37,6	0,191	30	Healthy

Рисунок 1 – Pima Indians Diabetes DataBase

В класифікаторі KNeighbors бібліотеки Sklearn є два важливих параметри: кількість сусідів і міра відстані між точками даних. Використання невеликої кількості сусідів працює добре, але оптимальне значення цього параметра можна підібрати для кожної ситуації, побудувавши графік точності при різних k. Щодо міри відстані між точками даних, найчастіше використовується Евклидова відстань, яка добре працює в багатьох ситуаціях. В класифікаторі ця відстань використовується за замовчуванням.

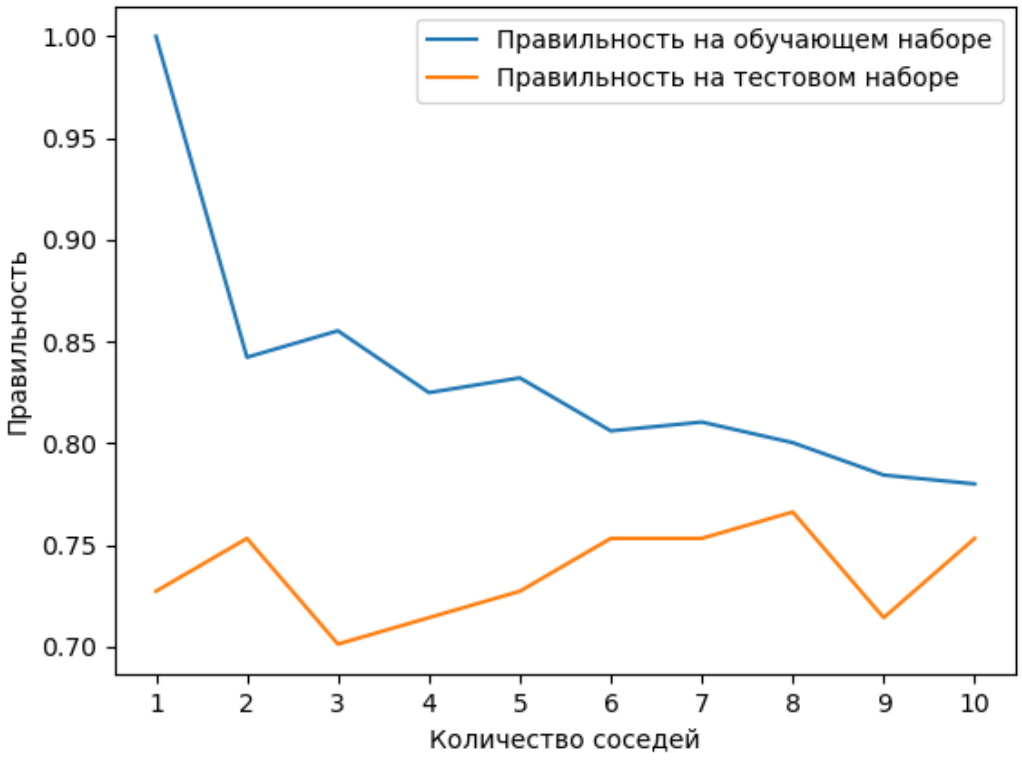

Рисунок 2 – Графік точності обчислення алгоритму при різних k

Точність класифікації (рис. 2) при такому розподілі даних складає 0.8.

Список використаної літератури

1. Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2017. – 418 с.: ил.

2. Мюллер, Андреас, Гвидо, Сара. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. : Пер. с англ. – СПб. : ООО "Альфа-книга", 2017. – 480 с. : ил. – Парал. тит. англ.

**Науковий керівник – Базілевич К. О., к.т.н., доц. каф. 304.*