Домашнее задание 7

Авласов Владислав

Задание 1

Код доступен в task1.py.
Скачиваем датасет с цветами. На стадии предобработки нормализуем данные и введём целочисленные значения для классов.
Будем использовать 5-fold validation, т.е. 90% данных – обучающая выборка, 10% – тестовая.
Датасет небольшой, поэтому можем перебрать большой диапазон нашего параметра k для соседей. Давайте переберём от 1 до 100.
Запустив несколько раз, можем увидеть, что из-за внесения рандома в алгоритм(мы шаффлим наши данные перед разделением) от запуска к запуску наблюдаются разные ответы.
Но! k в них не поднимается выше 25!
Приведу график одного из запусков. К сожалению, по непонятным мне причинам, мой компьютер не хочет ставить ни matplotlib, ни ggplot, поэтому график рисовал на стороннем сайте по точкам.
Синий график – ошибка на тесте, зелёный – на кросс-валидации.

Задание 2

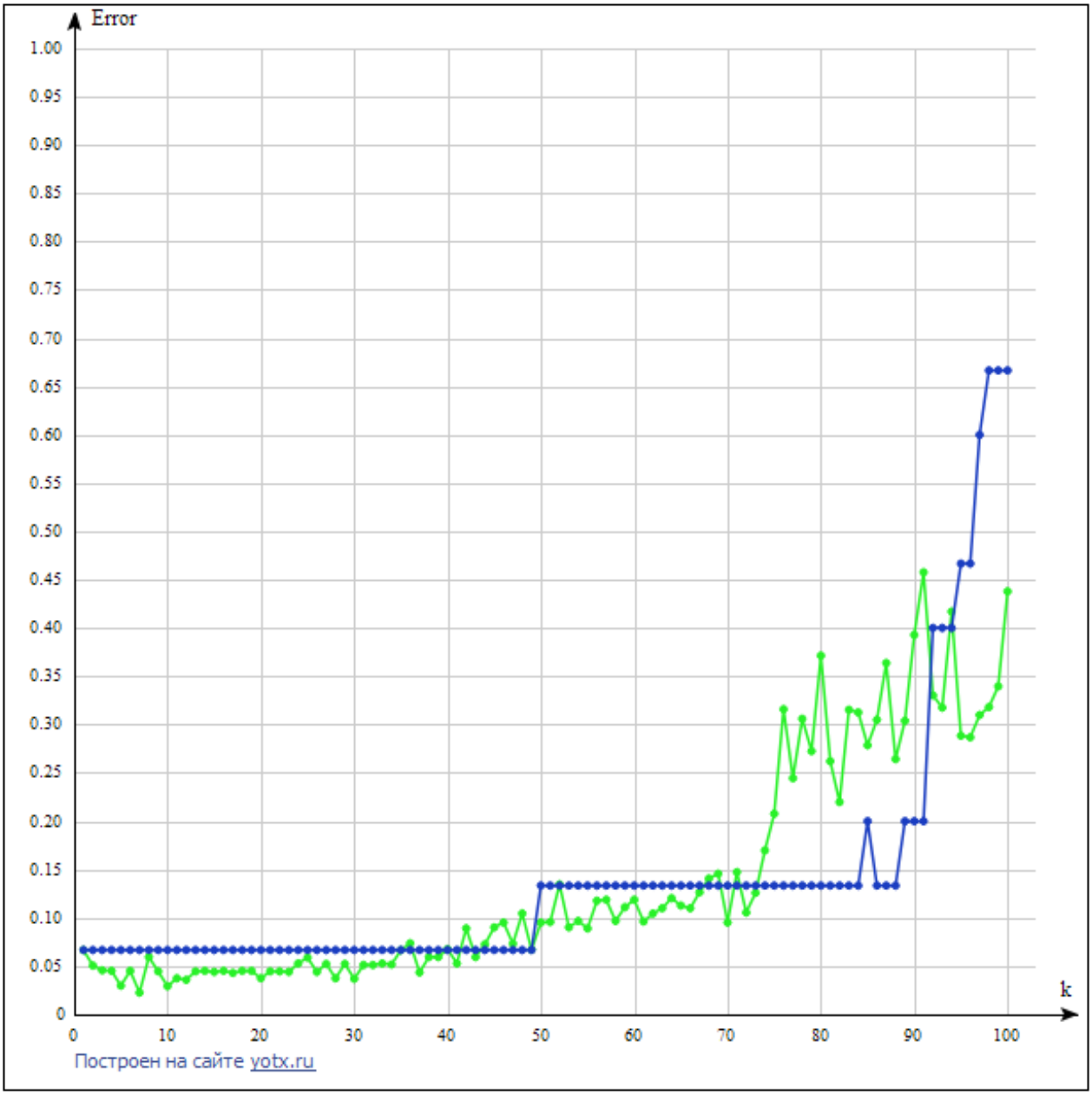

Рис. 1: Best result: k=13, CV accuracy = 97.09%, test accuracy = 93.33%