KNN Classifier

Authors: Krystian Ruszczak, Paweł Michalcewicz

Algorithm Overview

The k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method.

It does not build an explicit model during training. Instead it stores the training data and makes predictions only when new input samples are provided.

To classify new data, the algorithm looks at the k closest samples in the training set and assigns the class that appears most frequently among them. The closeness of two data points is measured using a distance metric - Euclidean distance.

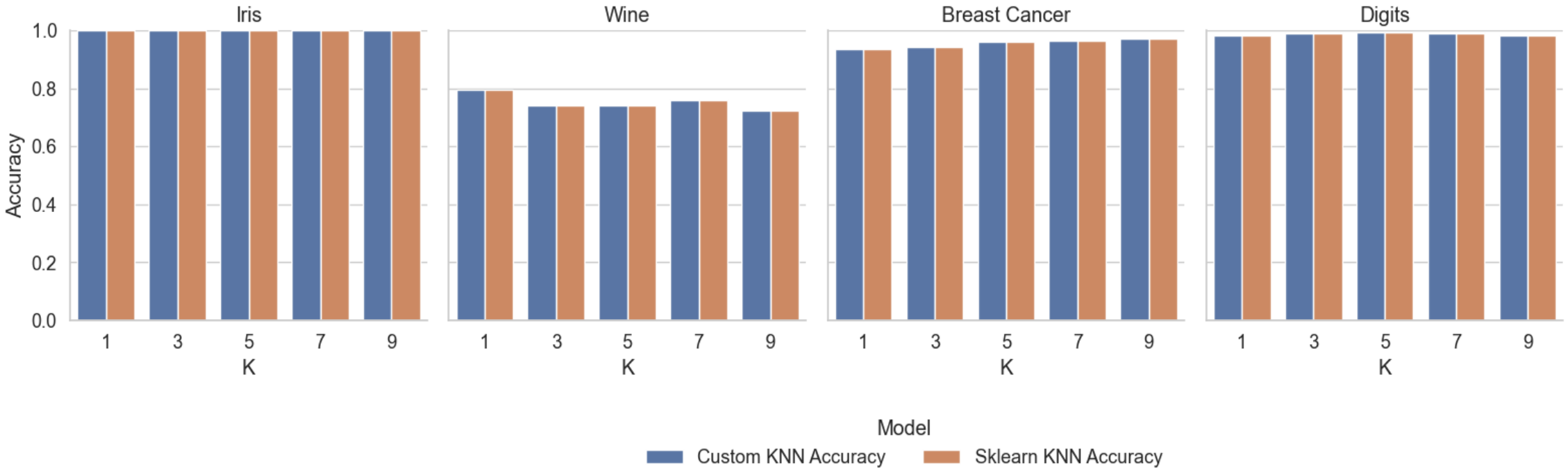
Implementation

The class KNNClassifier contains following methods:

- fit,
- _euclidean_distance,
- _predict_single,
- predict,
- score.

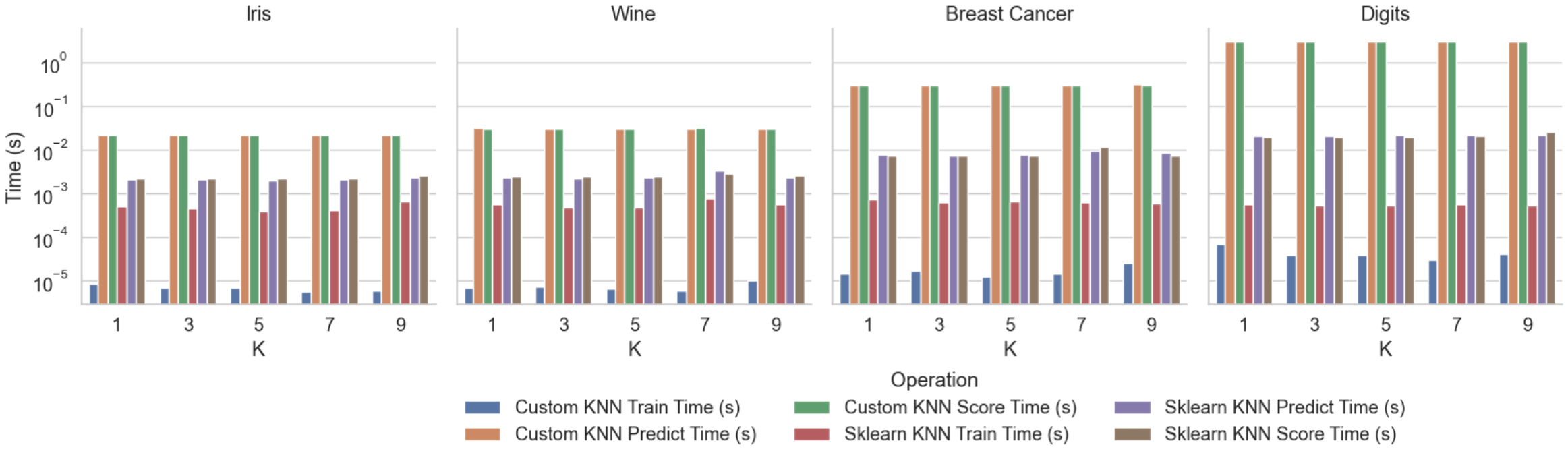
Comparison between two implementations

4 datasets were selected for the experiment: Iris, Wine, Breast Cancer, Digits. For each dataset, models were tested using five values of k: 1, 3, 5, 7 and 9.



Performance Comparison

Prediction and Scoring Time The most significant discrepancy appears in prediction speed: For small datasets (Iris, Wine), the custom KNN is slower than scikit-learn by roughly one order of magnitude. For medium datasets (Breast Cancer), the difference grows to 40 times slower. For the largest dataset (Digits), the difference becomes dramatic: Custom KNN prediction time 3.0 seconds, scikit-learn prediction time around 0.02 seconds.



Testing

The custom KNN model was checked using several small unit tests and one larger integration test. The unit tests made sure that the model starts with the right settings (test_initialization()), correctly stores the training data test_fit_storage(), calculates distances properly (test_euclidean_math()), and gives the expected predictions on simple examples (test_prediction_logic()). The final integration test compared the whole model to scikit-learn's KNN on the Iris dataset (test_compare_with_sklearn()). Both models returned exactly the same results, which confirms that the implementation works correctly and behaves just like the standard KNN classifier.

Thank You For Your Attention