Introduction to Machine Learning Program assignment #2

Abstract

Use K-NN classifier to analyze a data set.

Problem

Construct K-NN classifiers with

- Different algorithms
  - Linear Search [1]
  - Kd-Tree
  - Others, optional
- Different distance metrics
  - Manhattan distance
  - Euclidean distance
  - Cosine distance
  - Others, optional

Train these (at least) 6 models with data and compare their performance with computing time (for both model training and query time), confusion matrix, resubstitution validation and K-fold cross validation(K>2).

Submit your source code and report.

The report should include the results, environment, using library and language, explanation of your code and how to use it.

Data - White Wine Quality Data Set

Data Website: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

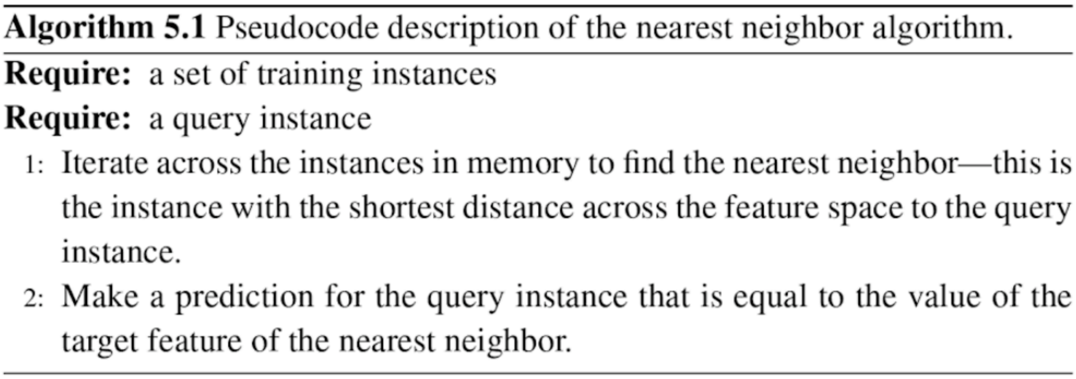
This data set contains 4898 instances with 12 attributes, where first 11 are feature inputs and the last one is the label (quality).

Use Data website to download **white wine** data set (winequality-white.csv) and get the full attribute information.

The data values are separated by ";" (semicolon), so you may want to convert them to "," which is used by standard CSV files.

Reference

[1]. Machine Learning for predictive data analytics, P.186.



Algorithm 5.1 Pseudocode description of the nearest neighbor algorithm.

```
Require: a set of training instances
Require: a query instance
Iterate across the instances in memory to find the nearest neighbor—this is
    the instance with the shortest distance across the feature space to the query
    instance.
Make a prediction for the query instance that is equal to the value of the
    target feature of the nearest neighbor.
```