ished in Towards Data Science

atsal
n 11, 2021 · 4 min read · Member-only · Listen

arest Neighbours Explained

tand the KNN algorithm and its implementation in Python
ıe sklearn library

Image from : https://unsplash.com/photos/lW25ZxpkIn8

rticle I will give a general overview, implementation, drawbacks and

s associated with the K Nearest Neighbours algorithm. Supervised is a subsection of machine learning generally associated with ation and regression based problems. Supervised learning implies are training a model using a labelled dataset. K Nearest Neighbours lls under the supervised learning umbrella and is one of the core ns in machine learning. It's a highly used, simple yet efficient of a non-parametric, lazy learner classification algorithm.

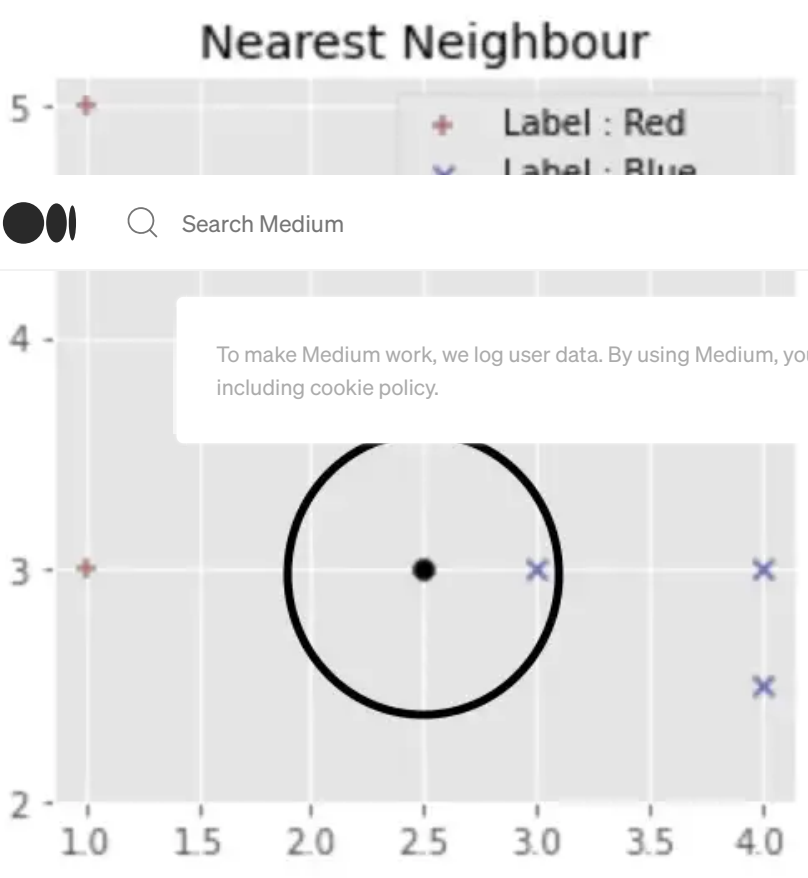
Learner implies that it doesn't learn a discriminative function from aining data but rather memorizes the training data instead

parametric implies that the algorithm makes no assumptions about istribution of the data.

N algorithm classifies unclassified data points based on their y and similarity to other available data points. The underlying ion this algorithm makes is that similar data points can be found another. It's commonly used to solve problems in various industries of its ease of use, application to classification and regression s, and the ease of interpretability of the results it generates.

orithm

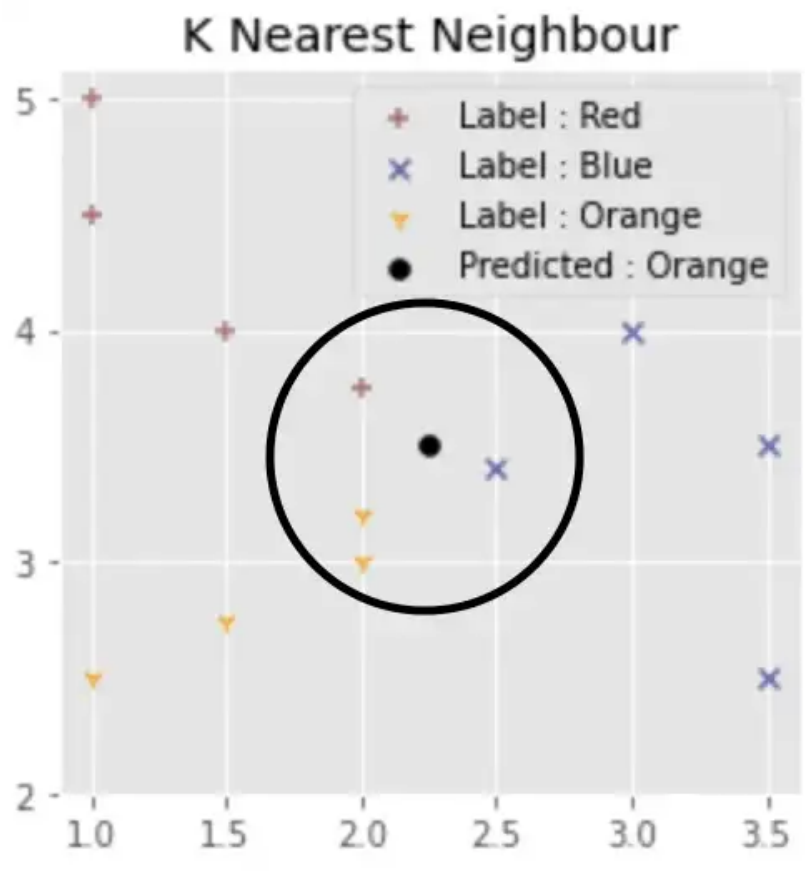
ents the number of nearest neighbours. When K = 1, the algorithm the nearest neighbour algorithm. This is the simplest scenario ven an unlabelled position X, the algorithm can predict its label by he closest labelled point to X and assigning that as the label.



enario, the unlabelled black point X would be predicted to be a blue point based on the nearest proximity of labelled points (Image provided by Author)

rithm works as follows :

```
ose the number of K and a distance metric used to calculate
oximity
d the K nearest neighbours of the point we want to classify
ign the point a label by majority vote
```



mber of orange points is larger than the number of blue and red points in the proximity of the black point. Thus the algorithm will predict this to be an orange label (Image provided by Author)

alue of K

ce of K is crucial for the model, if chosen incorrectly it can cause el to be over / under fit. A K value too small will cause noise in the ave a high influence on the prediction, however a K value too large e it computationally expensive.

istry standard for choosing the optimal value of K is by taking the oot of N, where N is the total number of samples. Of course, take this ain of salt as it varies from problem to problem.

experiment with various values of K and their associated accuracies. n practices to determine the accuracy of a KNN model is to use n matrices, cross validation or F1 scores.

ages & Disadvantages

ve listed some of the advantages and disadvantages of using the
orithm.

ges

le & intuitive — The algorithm is very easy to understand and
ement

ory based approach — Allows it to immediately adapt to new
ing data

ty of distance metrics — There is flexibility from the users side to
distance metric which is best suited for their application
idean, Minkowski, Manhattan distance etc.)

ntages

putational complexity — As your training data increases, the speed at
h calculations are made rapidly decrease

performance on imbalanced data — When majority of the data the
el is being trained on represents 1 label then that label will have a
likelihood of being predicted

nal value of K — If chosen incorrectly, the model will be under or
fitted to the data

entation

ıry

ıation, this article outlines that kNN is a lazy learner and non ric algorithm. It works by assigning a label to an unlabelled point ı the proximity of the unlabelled point to all the other nearest points. It's main disadvantages are that it is quite computationally nt and its difficult to pick the "correct" value of K. However, the ges of this algorithm is that it is versatile to different calculations of y, it's very intuitive and that it's a memory based approach.

ces

://scikit-
.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassi
ıtml#sklearn.neighbors.KNeighborsClassifier

enjoyed this read then check out my other works as well.

Chain Explained

rticle I will explain and provide the python implementations
ov chain. This article will not be a deep...

atascience.com

ediction Recommendation Engines with Node2Vec

ode Embeddings for Link Prediction

atascience.com

Vec Explained

ng the Intuition of Word2Vec & Implementing it in Python

atascience.com

nendation Systems Explained

ng & Implementing Content Based, Collaborative Filtering &
Recommendation Systems in Python

atascience.com

Carlo Method Explained

ost I will introduce, explain and implement the Monte Carlo
to you. This method of simulation is one of...

:om

Data Science Algorithms Machine Learning Knn Algorithm Python