IMPLEMENTATION OF KNN ALGORITHM

Nidhi Shetty, Srushti Bikkannavar, Shivani JA, Soumya Bhovi
KLS Gogte Institute of Technology - Belagavi

Abstract

- This analysis deals with exploring different morphometric features of cell nuclei and their relationship to diagnosis. The key morphometric features that we are going to focus on are the radius, perimeter, area, smoothness, compactness, concavity, and the number of concave points. Within each section of the analysis, we will explore each feature separately in relation to diagnosis (benign vs malignant) and then compare each feature to each other using the KNN Algorithm.

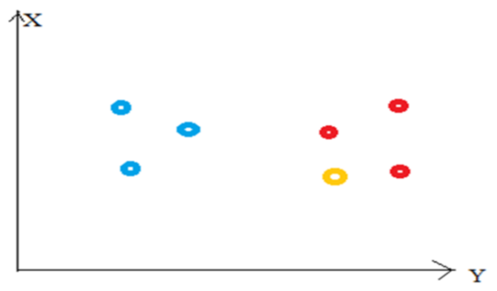
Introduction

- K Nearest Neighbor algorithm falls under the Supervised Learning category .
- It is used for classification (most commonly) and regression.
- It is a versatile algorithm also used for imputing missing values and resampling datasets.
- As the name (K Nearest Neighbor) suggests it considers K Nearest Neighbors (Data points) to predict the class or continuous value for the new Datapoint.

Algorithm

- Load the training data.
- Prepare data by scaling, missing value treatment, and dimensionality reduction as required.
- Find the optimal value for K:
- Predict a class value for new data:
    - ❑ Calculate distance(X, Xi) from i=1,2,3,….,n.
    - ❑ where X= new data point, Xi= training data, distance as per your chosen distance metric.
    - ❑ Sort these distances in increasing order with corresponding train data.
    - ❑ From this sorted list, select the top 'K' rows.
    - ❑ Find the most frequent class from these chosen 'K' rows. This will be your predicted class.

KNN Algorithm



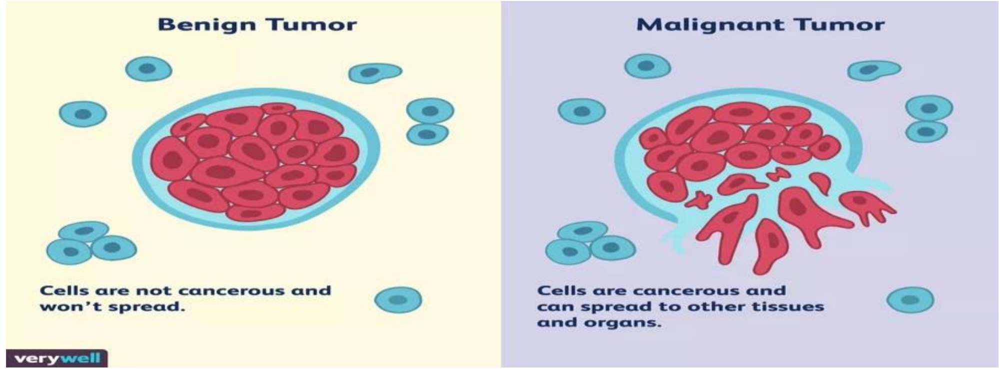
- Let us say we have plotted data points from our training set on a two-dimensional feature space.
- As shown, we have a total of 6 data points (3 red and 3 blue).
- Red data points belong to 'class1' and blue data points belong to 'class2'.
- And yellow data point in a feature space represents the new point for which a class is to be predicted.
- Obviously, we say it belongs to 'class1' (red points)
- Why?
- Because its nearest neighbors belong to that class!
-
- Yes, this is the principle behind K Nearest Neighbors. Here, nearest neighbors are those data points that have minimum distance in feature space from our new data point.
- And K is the number of such data points we consider in our implementation of the algorithm.
- Therefore, distance metric and K value are two important considerations while using the KNN algorithm.
- Euclidean distance is the most popular distance metric. You can also use Hamming distance, Manhattan distance, Minkowski distance as per your need.
- For predicting class/ continuous value for a new data point, it considers all the data points in the training dataset.
- Finds new data point's 'K' Nearest Neighbors (Data points) from feature space and their class labels or continuous values.
- For classification: A class label assigned to the majority of K Nearest Neighbors from the training dataset is considered as a predicted class for the new data point.
- For regression: Mean or median of continuous values assigned to K Nearest Neighbors from training dataset is a predicted continuous value for our new data point

IMPLEMENTATION OF ALGORITHM

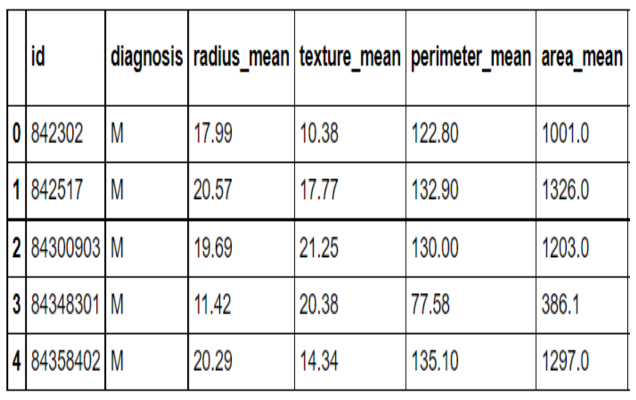
- We customized the data set into two sets training set and testing set by importing test_train_split from sklearn module.
- Then we dropped the columns with Nan and large integer values.
- We considered the diagnosis column as the target feature, with two attributes B-benign and M-malignant.
- So now we try to fit the training set data, fit function adjusts weights according to data values so that better accuracy can be achieved. After training, the model can be used for predictions.
- Then we import KNeighbhourClassifier through sklearn module to run the knn algorithm and try to fit the training set.
- Later, after training the test set will undergo testing through the algorithm in which later the accurarcy,confusion matrix and the classification report can be achieved.

Sample input/output

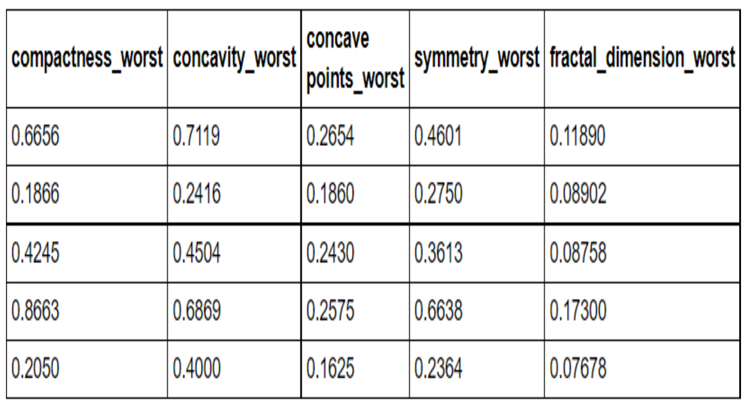
INPUT: We will use a Breast Cancer Analysis dataset used to determine whether the tumor is Benign or Malignant.


Benign Tumor
Cells are not cancerous and won't spread.
Malignant Tumor
Cells are cancerous and can spread to other tissues and organs.
verywell

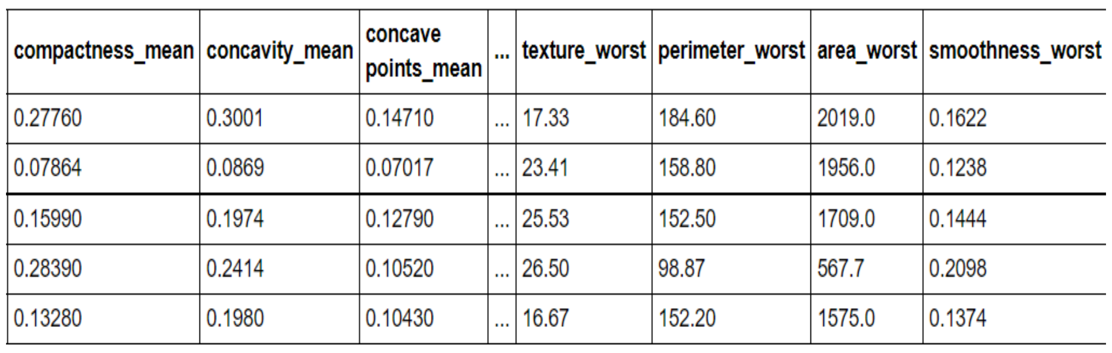
https://www.verywellhealth.com/what-does-malignant-and-benign-mean-514240



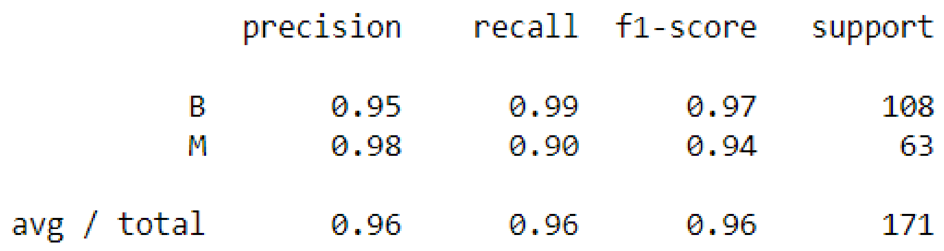
	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean
0	842302	M	17.99	10.38	122.80	1001.0
1	842517	M	20.57	17.77	132.90	1326.0
2	84300903	M	19.69	21.25	130.00	1203.0
3	84348301	M	11.42	20.38	77.58	386.1
4	84358402	M	20.29	14.34	135.10	1297.0



compactness_worst	concavity_worst	concave points_worst	symmetry_worst	fractal_dimension_worst
0.6656	0.7119	0.2654	0.4601	0.11890
0.1866	0.2416	0.1860	0.2750	0.08902
0.4245	0.4504	0.2430	0.3613	0.08758
0.8663	0.6869	0.2575	0.6638	0.17300
0.2050	0.4000	0.1625	0.2364	0.07678



compactness_mean	concavity_mean	concave points_mean	...	texture_worst	perimeter_worst	area_worst	smoothness_worst
0.27760	0.3001	0.14710	...	17.33	184.60	2019.0	0.1622
0.07864	0.0869	0.07017	...	23.41	158.80	1956.0	0.1238
0.15990	0.1974	0.12790	...	25.53	152.50	1709.0	0.1444
0.28390	0.2414	0.10520	...	26.50	98.87	567.7	0.2098
0.13280	0.1980	0.10430	...	16.67	152.20	1575.0	0.1374



```
             precision    recall  f1-score   support

          B       0.95      0.99      0.97       108
          M       0.98      0.90      0.94        63

avg / total       0.96      0.96      0.96       171
```

CONCLUSION

By using knn algorithm we were successfully able to predict whether the tumor was benign or malignant and the accuracy was around 95% .
We have successfully achieved this algorithm with the help of module's that are available in the python.

REFERENCES

- https://www.analyticsvidhya.com/blog/2021/04/simple-understanding-and-implementation-of-knn-algorithm/
- https://www.kaggle.com/niteshyadav3103/breast-cancer-classification