Supervised Learning Classification

Agenda

- KNN Basics
- KNN algorithm

K - Nearest Neighbours

Data matrix

- The data represented in form of a matrix is called the data matrix
- For data with m features are n observations the data matrix is given as

$$\begin{pmatrix} x_{11} & x_{21} & \cdots & x_{m1} \\ x_{12} & x_{22} & \cdots & x_{m2} \\ & & \ddots & \\ x_{1n} & x_{2n} & \cdots & x_{mn} \end{pmatrix}$$

Proximity Measures

Proximity measures

- The proximity measures find the distance between two instances
- Proximity measures include
  - Similarity measures
  - Dissimilarity measure
- Depending upon the data types, we choose the proximity measure

Similarity measures

A similarity measure for two objects, will return the value 0 if the objects are unlike, and the value 1 if the objects are alike.

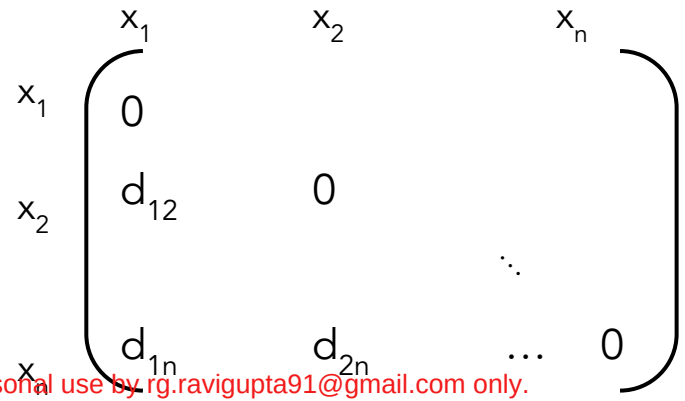
A similarity matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ \\ x_n \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & & x_n \end{array} \\ \begin{pmatrix} 1 & & & \\ d_{12} & 1 & & \\ & & \ddots & \\ d_{1n} & d_{2n} & \cdots & 1 \end{pmatrix} \end{array}$$

Dissimilarity measures

- Dissimilarity measure work exactly opposite of a similarity measure
- A dissimilarity measure for two objects, will return the value 1 if the objects are unlike, and the value 0 if the objects are alike

A dissimilarity matrix



Distance measures

Depending on the type of data we have different distance measures

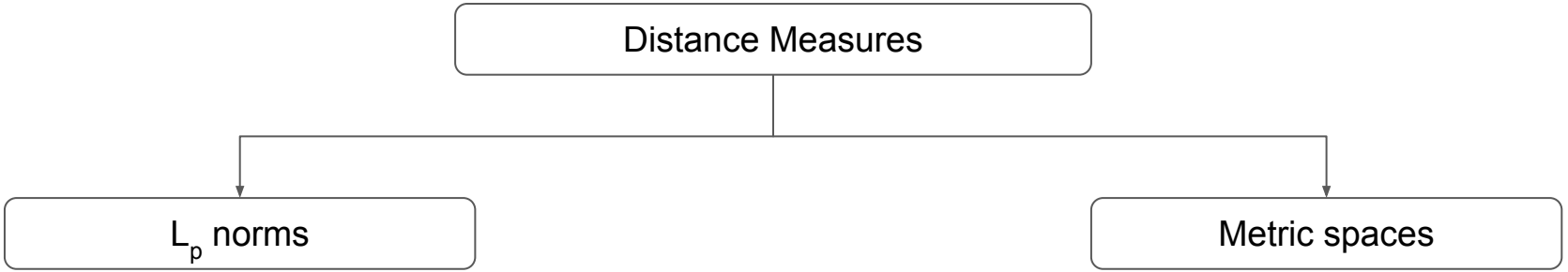
Numeric data

- Euclidean distance
- Manhattan distance
- Minkowski distance
- Chebyshev's distance

String data

- Cosine distance
- Edit distance
- Longest Common Sequence
- Hamming distance

Distance measures



The distance measures based on norm

The distance measures which satisfy the following properties:

$d(a, b) \geq 0$ (non-negativity)

$d(a, b) \equiv 0 \iff a \equiv b$ (positive definiteness)

$d(a, b) \equiv d(b, a)$ (symmetry)

$d(a, b) \leq d(a, c) + d(c, b)$ (triangle inequality)

Euclidean distance - numeric data

- Euclidean distance is obtained for numeric data
- It is the L2 norm
- For two instances X and Y the Euclidean distance is given by

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where x_i and y_i are the values taken by X and Y respectively

- More the distance between the two instance more the dissimilarity measure

Euclidean distance - numeric data

Example: Consider two points (5,6) and (1, 3). Obtain the Euclidean distance.

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(5-1)^2 + (6-3)^2}$$
$$= \sqrt{16+9}$$
$$= \sqrt{25}$$
$$= 5$$

Squared euclidean distance - numeric data

- Similar to Euclidean distance is obtained for numeric data
- Just the squared value of Euclidean distance
- For two instances X and Y the Euclidean distance is given by

$$\sum_{i=1}^{n} (x_i - y_i)^2$$

where x_i and y_i are the values taken by X and Y respectively

Manhattan distance - numeric data

- It is the L1 norm
- For two instances X and Y, it is given by

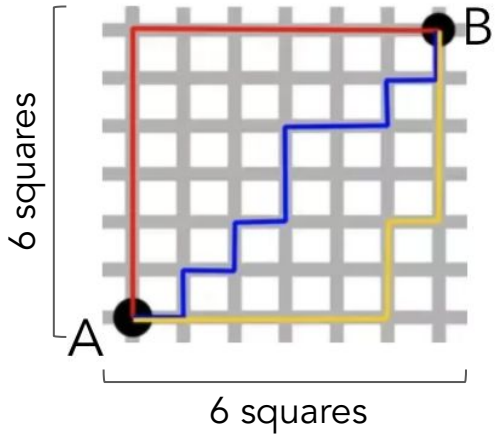
$$\sum_{i=1}^{n} |x_i - y_i|$$

where x_i and y_i are the values taken by X and Y respectively

- Also known as the Taxicab distance or Snake distance or the City block distance

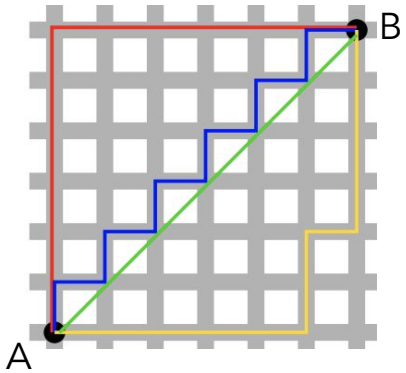
Manhattan distance - numeric data

- Consider points A and B
- The Manhattan distance between the points is given by the edge of the squares it crosses
- Here the Manhattan distance is 12



Manhattan distance - numeric data

- For the same points the Euclidean distance is the given by the shortest distance between them
- Given by the green line
- The euclidean distance is $6\sqrt{2}$



Minkowski distance - numeric data

- It is the generalized form of the Manhattan and the Euclidean distance
- It is the Lp norm
- For two instances X and Y, it is given by

$$\sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

where x_i and y_i are the values taken by X and Y respectively and p>0

Minkowski distance - numeric data

Example: Consider two points (5,6) and (1, 3). Obtain the minkowski distance (take p = 4)

$$\sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p} = \sqrt[4]{\sum_{i=1}^{n}(x_i - y_i)^4}$$

$$= \sqrt[4]{(5-1)^4 + (6-3)^4}$$

$$= \sqrt[4]{256 + 81} = \sqrt[4]{337}$$

$$= 4.28$$

Chebyshev's distance - numeric data

- It is the L∞ norm
- For two instances X and Y

$$\max_{i=1}^{n} |x_i - y_i|$$

where x_i and y_i are the values taken by X and Y respectively

Chebyshev's distance - numeric data

Example: Consider two points (5,6) and (1, 3). Obtain the Chebyshev's distance.

$$\max_{i=1}^{n} |x_i - y_i| = \max\{|5-1|, |6-3|\}$$

$$= \max\{4, 3\}$$

$$= 4$$

K - NN algorithm

K - NN algorithm

Specifies number of nearest neighbours

NN stands for Nearest Neighbours

KNN algorithm

- The K - Nearest Neighbour (KNN) algorithm classifies the data based on the similarity measure
- K specifies the number of nearest neighbours to be considered
- Does not require the data to be trained

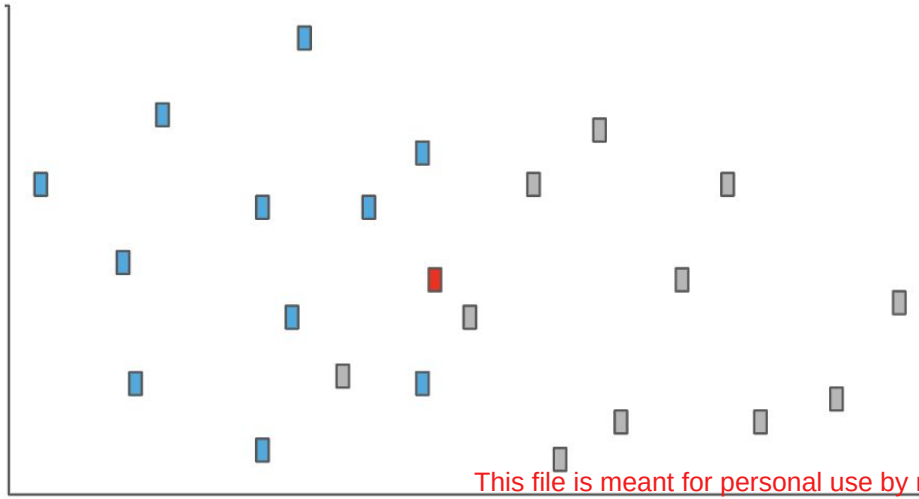
KNN algorithm

KNN is considered to be:

- Instance based learning algorithm: uses training instances to make predictions
- Lazy learning algorithm: does not required a model to be trained
- Non-Parametric algorithm: no assumptions are made about the functional form of the the problem being solved

KNN algorithm

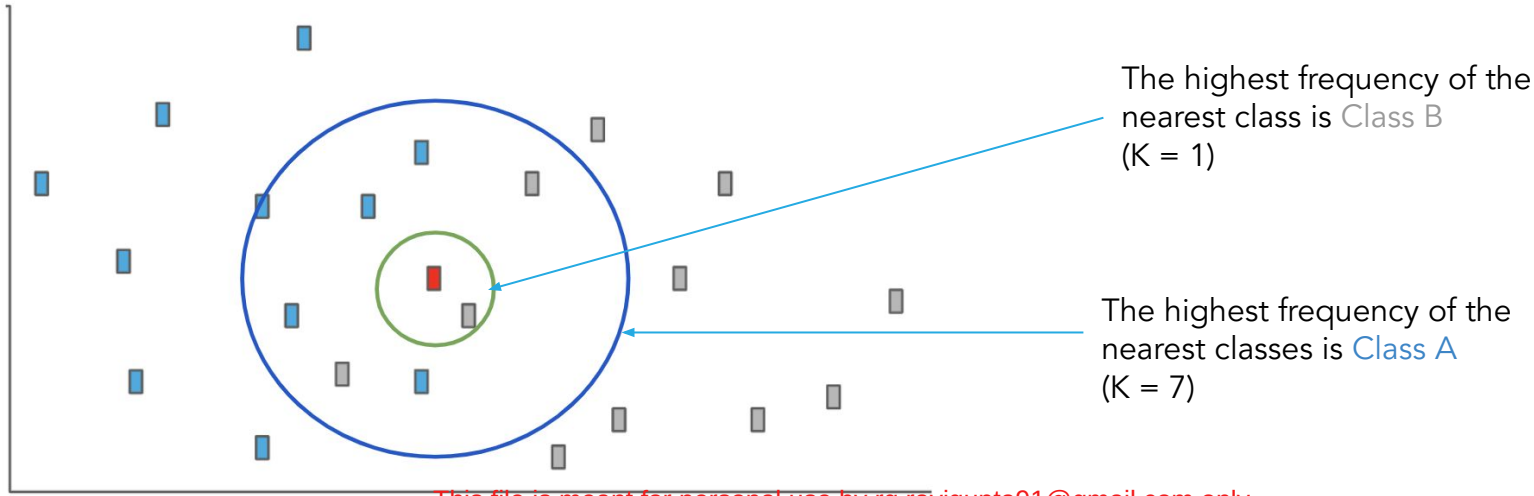
Consider two classes as shown in the figure - Class A and Class B



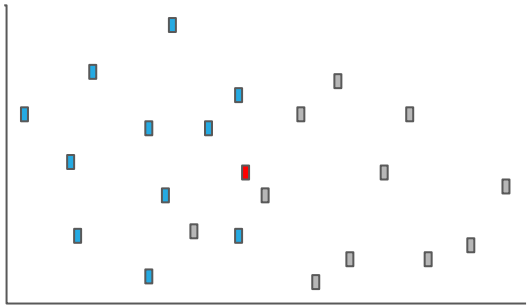
Which class does the red point belong to?

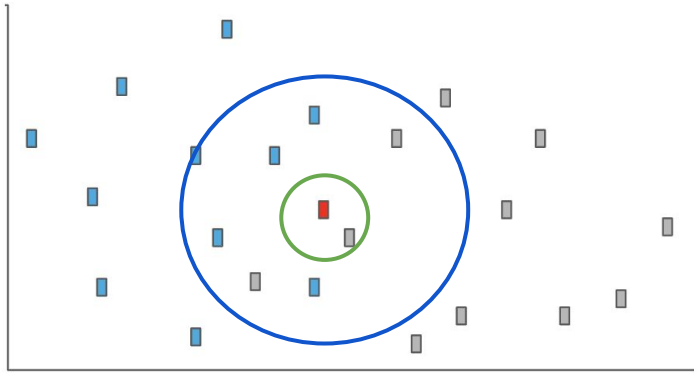
KNN algorithm

An easy approach is to label it as the class which has the highest frequency around it

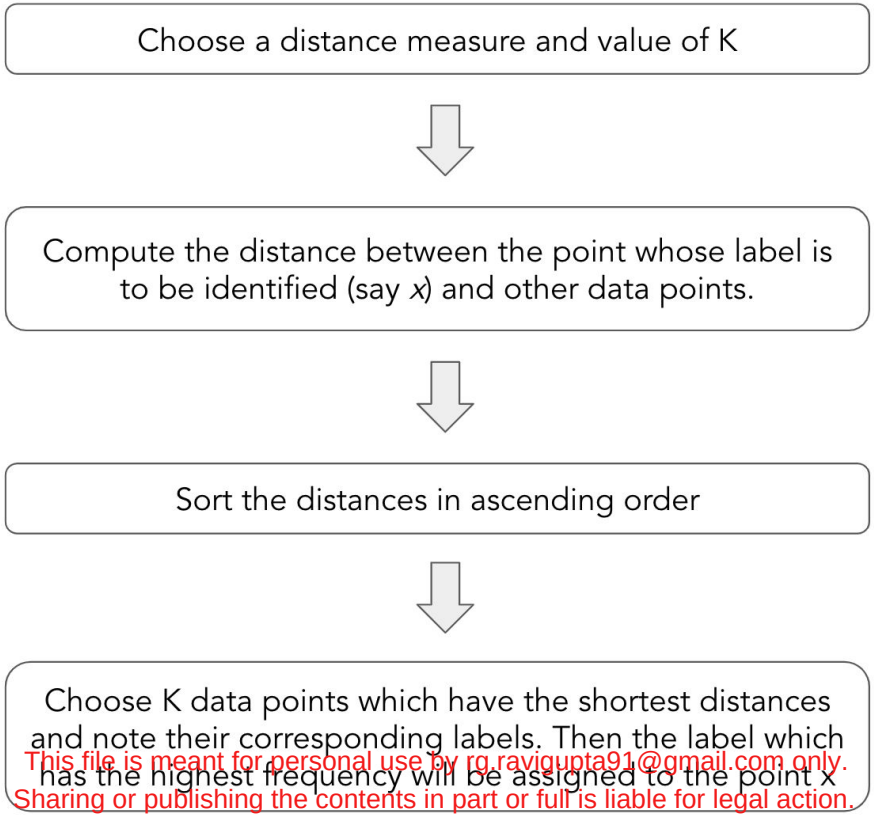


KNN algorithm





KNN algorithm - Procedure



Choose a distance measure and value of K

Compute the distance between the point whose label is to be identified (say *x*) and other data points.

Sort the distances in ascending order

Choose K data points which have the shortest distances and note their corresponding labels. Then the label which has the highest frequency will be assigned to the point x

Normalization of data

In order to perform KNN, the data needs to be normalized.

Consider a data with 5 features, of which 4 features have small range (say between 0 to 25) and the fifth feature ranges from -100 to 2000. The classification would be majorly based on the feature with high range. Since it's contribution to the distance measure would be high with very little or no effect of the other variables thus we normalise the data.

Example

- Humidity: (Independent variable) the percentage of humidity in the atmosphere

- Temperature: (Independent variable) the temperature average temperature during precipitation

- Rain: (Target variable) indicates whether it rained or not; takes value 1 if rained and value 0 otherwise

Observation	Humidity	Temperature	Rain
1	58	19	0
2	62	26	0
3	40	30	0
4	36	35	0
5	87	19	1
6	93	18	1
7	79	16	1
8	69	17	1
9	62	33	0
10	71	15	1
11	55	33	0
12	78	19	1
13	60	20	1
14	58	35	0
15	35	39	0

Would it rain if Humidity = 84 and Temperature = 37?

Example

- Let us choose the K = 5 and used the Euclidean distance
- Compute the Euclidean distance between the new data for each instance
- For example, consider the first observation Humidity = 58 and Temperature = 19, the Euclidean distance is

$$[(58 - 84)^2+(19 -34)^2]^{1/2} = 31.623$$

Humidity	Temperature	Rainfall
84	34	?

Example

Computed the Euclidean distances for each instance with the new data and sort the data in ascending order with respect to the Euclidean distance

Observation	Euclidean Distance (sorted)	Class Label (Rainfall)
5	18.25	1
12	18.97	1
6	21.02	1
7	21.59	1
9	22.36	0
2	24.6	0
8	25.00	1
10	25.55	1
14	26.08	0
11	29.27	0
13	29.41	1
1	31.62	0
3	44.55	0
4	48.04	0
15	49.04	0

Example

- Since K = 5 consider the class labels of first five observations
- 1 appears 4 times and 0 appears 1 time

Observation	Euclidean Distance (sorted)	Class Label (Rainfall)
5	18.25	1
12	18.97	1
6	21.02	1
7	21.59	1
9	22.36	0
2	24.6	0
8	25.00	1
10	25.55	1
14	26.08	0
11	29.27	0
13	29.41	1
1	31.62	0
3	44.55	0
4	48.04	0
15	49.04	0

Example

- 1 appears 4 times and 0 appears 1 time
- Thus for our new instance the class label is 1 (using max voting)

Humidity	Temperature	Rainfall
84	34	1

- Implies that for Humidity = 84 and Temperature = 34, it will rain

The value of K

In order to avoid a tie, while using KNN:

- For even number of class labels consider K to be odd
- For odd number of class labels consider K to be even

The value of K

The value of K should be chosen appropriately since a large K value may reduce the variance due to the noisy data, but increase bias resulting to ignorance of smaller patterns, whereas a small K may overfit the data

Weighted KNN

- Selecting an apt K is challenging. To overcome this, weighted KNN is used
- Weights are assigned to each instance
- Generally the weights are the inverse of the the distance
- The weights are higher for the points which are nearer to the new instance
- The weights are lower for the points which are away from the new instance

KNN

Advantages

- Easy to implement
- No training required
- New data can be added at any time
- Effective if training data is large

Disadvantages

- To chose apt value for K
- Computational expensive
- Can not tell which features gives the best result

KNN - Applications

- Image classification
- Handwriting recognition
- Predict credit rating of customers
- Replace missing values

Thank You