Machine Learning

K-Nearest Neighbor

Satishkumar L. Varma

Department of Information Technology
SVKM's Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai.
ORCID | Scopus | Google Scholar | Google Site | Website

Outline

- Classification
  - Bayesian Belief Networks
  - Hidden Markov Models
  - Support Vector Machine
    - Maximum Margin Linear Separators
    - Quadratic Programming solution to finding maximum margin separators
    - Kernels for learning non-linear functions
  - Classification using k Nearest Neighbour Algorithm

K-Nearest Neighbor

- K-Nearest Neighbour (KNN)
- It is one of the Supervised learning algorithm
- It is mostly used for classification of data on the basis how it's neighbour are classified.
- It stores all available cases and classifies new cases based on a similarity measure.
- K in KNN is a parameter that refers to the number of the nearest neighbours
- K nearest neighbours are used to find the majority voting for identify the class.
- It is also called a "Lazy learner".

K-Nearest Neighbor

- Selection of value of K
  - Sqrt(n)
  - where n is a total number of data points
  - If in case n is even
    - we have to make the value odd by adding 1 or subtracting 1 that helps in select better
- When to use KNN
- Dataset is labelled
- Dataset is noise-free
- Dataset is small because it is a “Lazy learner”

K-Nearest Neighbor

- The KNN classification algorithm

$$y' = \underset{v}{\operatorname{argmax}} \sum_{(x_i, y_i) \in D_z} I(v = y_i)$$

- Let k be the number of nearest neighbors and D be the set of training exampıes.
- 1. for each test example z = (x', y') do
- 2. Compute d(x', x), the distance between z and every example, (x, y) ϵ D
- 3. Select Dz ⊆ D, the set of k closest training examples to z.
- 4. Compute y' = argmax$v \sum I(v = yi)$ for $\boldsymbol{x}i,yi \in Dz$
- 5. End for

K-Nearest Neighbor

- KNN is a "Lazy learner"
- Example: K-Nearest Neighbour

Player	Age	Gender	Class
A	32	0	Football
B	40	0	Neither
C	16	1	Cricket
D	34	1	Cricket
E	55	0	Neither
F	40	0	Cricket
G	20	1	Neither
H	15	0	Cricket
I	55	1	Football
J	15	0	Football

Note: Here male is denoted with numeric value 0 and female with 1.

Question: Find in which class of sports person X lie whose k factor is 3 and age is 5.

Euclidean distance $d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$

Manhatten distance $d(p,q) = \sum_i |p_i - q_i|$

q norm distance $d(p,q) = (\sum_i |p_i - q_i|^q)^{1/q}$

K-Nearest Neighbor

- Example: K-Nearest Neighbour
- To find the distance (d) between any two points using say Euclidean Distance:
- $d=\sqrt{((x2-x1)^2+(y2-y1)^2)}$
- To find out the distance between A and X as follows:
- $d=\sqrt{((age2-age1)^2+(gender2 - gender1)^2)}$
- $d=\sqrt{((5-32)^2+(1-0)^2)}$
- $d=\sqrt{729+1}$
- $d=27.02$

K-Nearest Neighbor

- Example: K-Nearest Neighbour
- To find the d between any two points using say Euclidean Distance:

Euclidean distance $d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$

Manhatten distance $d(p,q) = \sum_i |p_i - q_i|$

q norm distance $d(p,q) = (\sum_i |p_i - q_i|^q)^{1/q}$

Similarly, we find out all distance one by one.

D (X & ?)	Distance (d)	Class
A	27.02	Football
B	35.01	Neither
C	11	Cricket
D	29	Cricket
E	50.01	Neither
F	35.01	Cricket
G	15	Neither
H	10.05	Cricket
I	50	Football
J	10.05	Football

D (X & ?)	Sorted d	Class
H	10.05	Cricket
J	10.05	Football
C	11	Cricket
G	15	Neither
A	27.02	Football
D	29	Cricket
B	35.01	Neither
F	35.01	Cricket
I	50	Football
E	50.01	Neither

K-Nearest Neighbor

- Example: K-Nearest Neighbour
- As the value of k=3 for person X;
- The first K=3 closest person (as highlighted with blue) are
- H: 10.05 Cricket; J 10.05 Football; and C 11 Cricket
- And the voting majority is Cricket so person X is classified as Cricket.
- i.e as per KNN algorithm; the person X will be in the class of people who like cricket.

Euclidean distance $d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$

Manhatten distance $d(p,q) = \sum_i |p_i - q_i|$

q norm distance $d(p,q) = (\sum_i |p_i - q_i|^q)^{1/q}$

References

Text books:

1. Ethem Alpaydin, “Introduction to Machine Learning , 4th Edition, The MIT Press, 2020.
2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012."
3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.
4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1ed, O'reilly, 2016.
5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012."

Reference Books:

6. Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, 2nd Edition, Shroff/O'Reilly, 2019.
7. Witten Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal., “Data Mining: Practical machine learning tools and techniques”, 1st Edition, Morgan Kaufmann, 2016.
8. Han, Kamber, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012.
9. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, “Foundations of Machine Learning”, 1ed, MIT Press, 2012.
10. H. Dunham, “Data Mining: Introductory and Advanced Topics”, 1st Edition, Pearson Education, 2006.

Thank You.