k Nearest Neighbors

Thanh-Nghi Do

Can Tho University

dtnghi@cit.ctu.edu.vn

Can Tho
Dec. 2019

Content

- Introduction
- *k* Nearest Neighbors

Content

- **<u>Introduction</u>**
- *k* Nearest Neighbors

Training dataset => classification

X1	X2	Class
0	0	+1
0	1	+1
0	2	+1
1	1	+1
2	0	-1
2	1	-1
3	1	-1

X1	X2	Class
0	2	?

Training dataset => classification

X1	X2	Class
0	0	+1
0	1	+1
0	2	+1
1	1	+1
2	0	-1
2	1	-1
3	1	-1

X1	X2	Class
0	2	?

⇨ **class = +1**

Training dataset => classification

X1	X2	Class
0	0	+1
0	1	+1
0	2	+1
1	1	+1
2	0	-1
2	1	-1
3	1	-1

X1	X2	Class
2	1	?

Training dataset => classification

X1	X2	Class
0	0	+1
0	1	+1
0	2	+1
1	1	+1
2	0	-1
2	1	-1
3	1	-1

X1	X2	Class
2	1	?

⇨ **class = -1**

Training dataset => classification

X1	X2	Class
0	0	+1
0	1	+1
0	2	+1
1	1	+1
2	0	-1
2	1	-1
3	1	-1

X1	X2	Class
1.1	1	?

Training dataset => classification

X1	X2	Class
0	0	+1
0	1	+1
0	2	+1
1	1	+1
2	0	-1
2	1	-1
3	1	-1

X1	X2	Class
1.1	1	?

⇒ **class = +1**

Top 10 Data Mining Algorithms (Kdnuggets)

Here are the algorithms:

- 1. C4.5
- 2. k-means
- 3. Support vector machines
- 4. Apriori
- 5. EM
- 6. PageRank
- 7. AdaBoost
- 8. kNN
- 9. Naive Bayes
- 10. CART

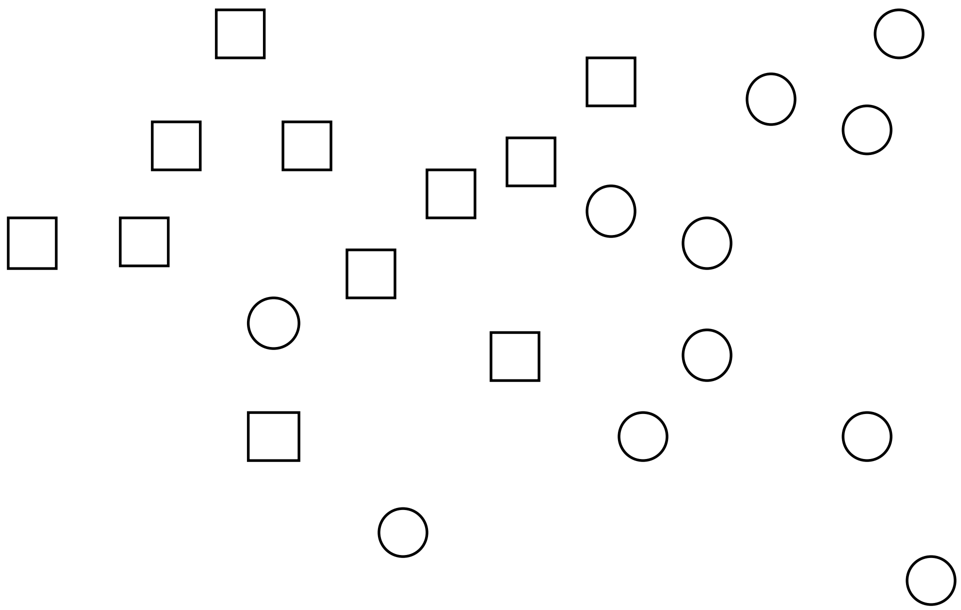
k Nearest Neighbors (*k*NN)

- *k* Nearest Neighbors (*k*NN)
  - Learning by analogy
  - Simple and intuitive
  - Called lazy learning (no training step)
  - Training examples themselves represent the knowledge
  - New example x is classified into the label which is most frequent among the k training examples nearest to that point x
  - Dependent on the distance measure
  - Can be extended for regression and clustering

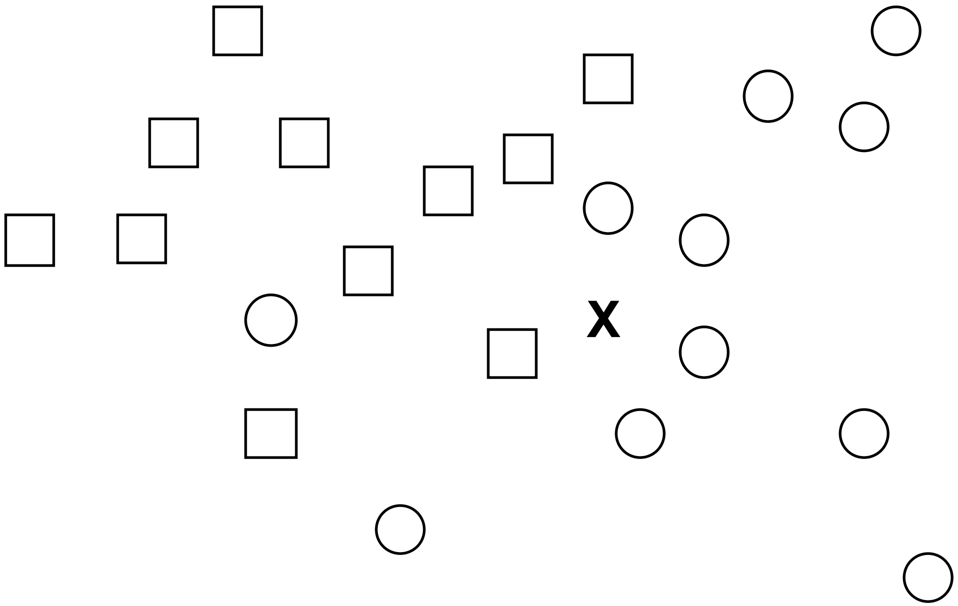
Content

- Introduction
- ***k* Nearest Neighbors**

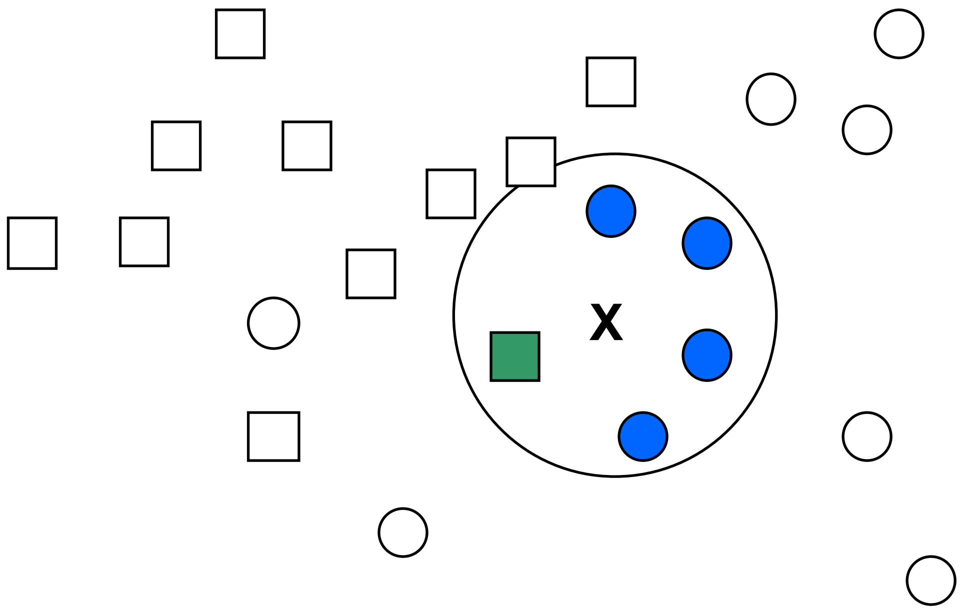
k Nearest Neighbors (*k*NN)



k Nearest Neighbors (*k*NN)



k Nearest Neighbors (*k*NN)



Distance measures

- Distance measure between x, y: $d(x, y)$
- Verifying 4 axioms for all x, y, z
  1. Non negativity: $d(x, y) \geq 0$
  2. Zero property: $d(x, x) = 0$
  3. Symmetry: $d(x, y) = d(y, x)$
  4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

Distance measures

- Minkowski distance

$$d(u,v)=\sqrt[q]{(|u_1-v_1|^q+|u_2-v_2|^q+\ldots+|u_n-v_n|^q)}$$

two points u, v in n-dimensional input space,

$u = (u_1, u_2, \ldots, u_n)$ *and* $v = (v_1, v_2, \ldots, v_n)$

$q >= 1$

Distance measures

- Minkowski distance
- with *q = 1* => *d* is Manhattan distance

$$d(u,v) = |u_1 - v_1| + |u_2 - v_2| + \ldots + |u_n - v_n|$$

- with *q = 2* => *d* is Euclidean distance

$$d(u,v) = \sqrt{(|u_1 - v_1|^2 + |u_2 - v_2|^2 + \ldots + |u_n - v_n|^2)}$$

Example

X1	X2	Class
0.45	5	?

X1	X2	Class	d(Manhattan)
0.1	10	+1	5.35
0.2	25	+1	20.25
0.3	0	+1	5.15
0.5	11	-1	6.05
0.8	100	-1	95.35
0	50	+1	45.45
1	70	-1	65.55

1NN ⇒ **Class = +1**

Comments

- Feature *X2* in *[0..100]*
- Feature *X1* in *[0..1]*
- Distance measure depends on *X2*
- Necessary to normalize data
- Scaling *X2* to have values between *0* and *1*

$$new\_val = (val - min)/(max - min)$$

Normalized *X2*

X1	X2	Class
0.45	0.05	?

X1	X2	Class
0.1	0.1	+1
0.2	0.25	+1
0.3	0	+1
0.5	0.11	-1
0.8	1	-1
0	0.5	+1
1	0.7	-1

D(Manhattan)
0.4
0.45
0.2
0.11
1.3
0.9
1.2

1NN ⇨ **class = -1**

Discussion

- Simple and intuitive: *k* nearest neighbors
- Tuning *k*
- Choosing the distance measure
- No training step
- Training examples themselves represent the knowledge
- Very slow for the prediction: simple version scans entire training example to derive a prediction
- Statisticians have used *k*NN since early 1950s

  If $n \to \infty$ and $k/n \to 0$, error approaches minimum

Discussion

- Theorem: For sufficiently large training set size *n*, the error rate of the *1*NN classifier is less than twice the Bayes error rate. (Cover & Hart, 1967)

Merci !