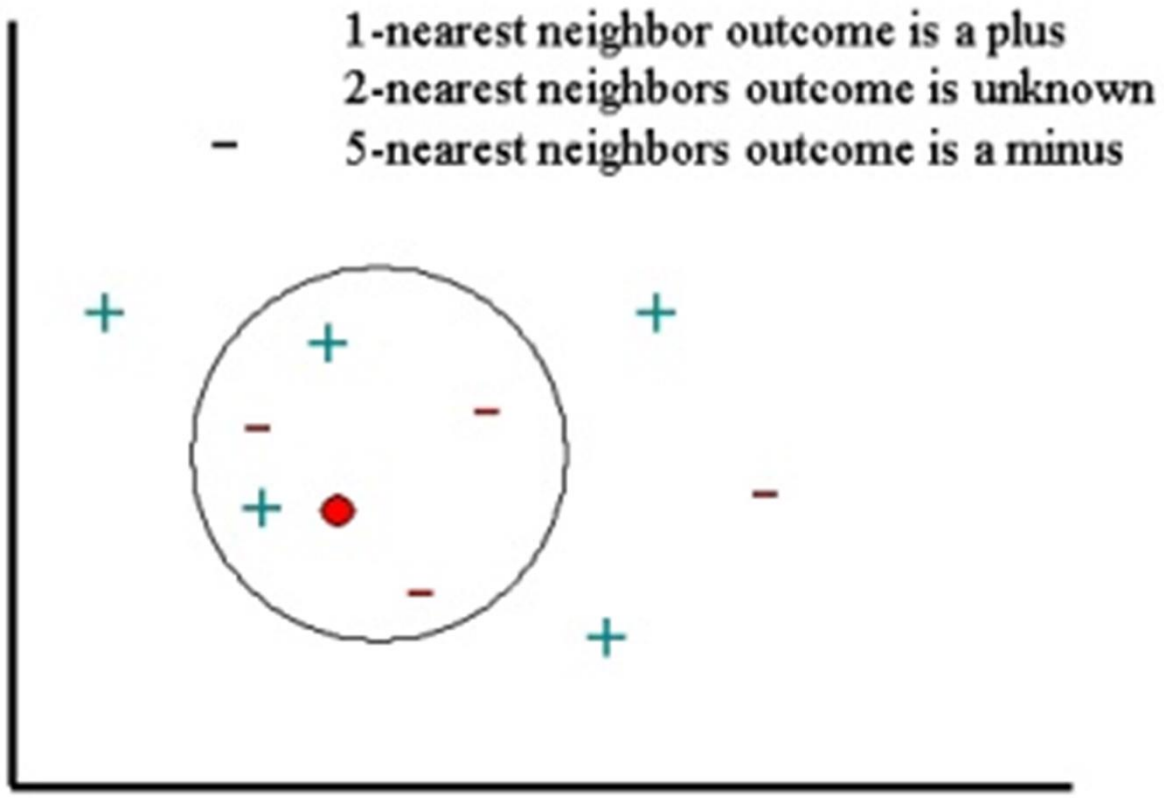
K NEAREST NEIGHBOUR

Abdus Salam Azad

K Nearest Neighbor



K Nearest Neighbor - Parameters

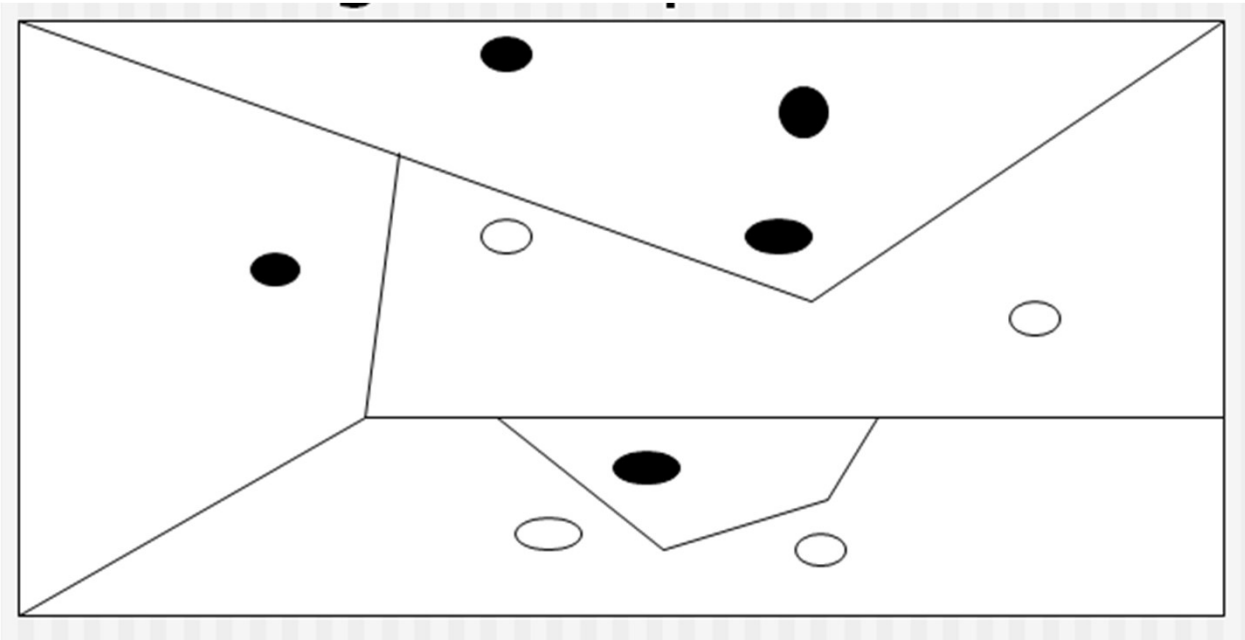
- Value of k
- Distance measurement
- Finding the neighbor
  - Linear search, kdtree

kNN

- Lazy Learner
  - do no real work until classification time

Decision Boundary

- the decision surface induced by 1-NN is a voronoi diagram



Pros and Cons

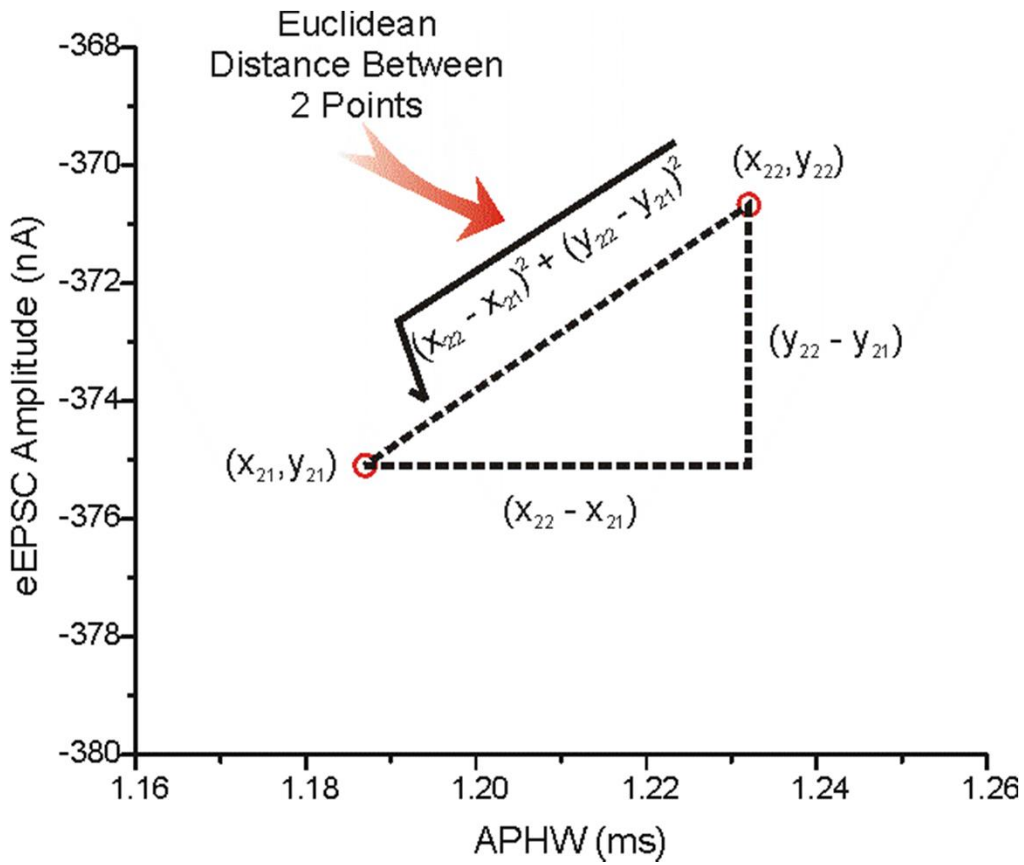
- + Effective inductive inference method for complex target functions
- +Learning is very simple
- - Classification is time consuming

Attribute Types

- Numeric
  - Continuous, Discrete
- Nominal
  - Yellow, Blue, White
- Ordinal
  - Small, Medium, Large
- Boolean
  - True,False

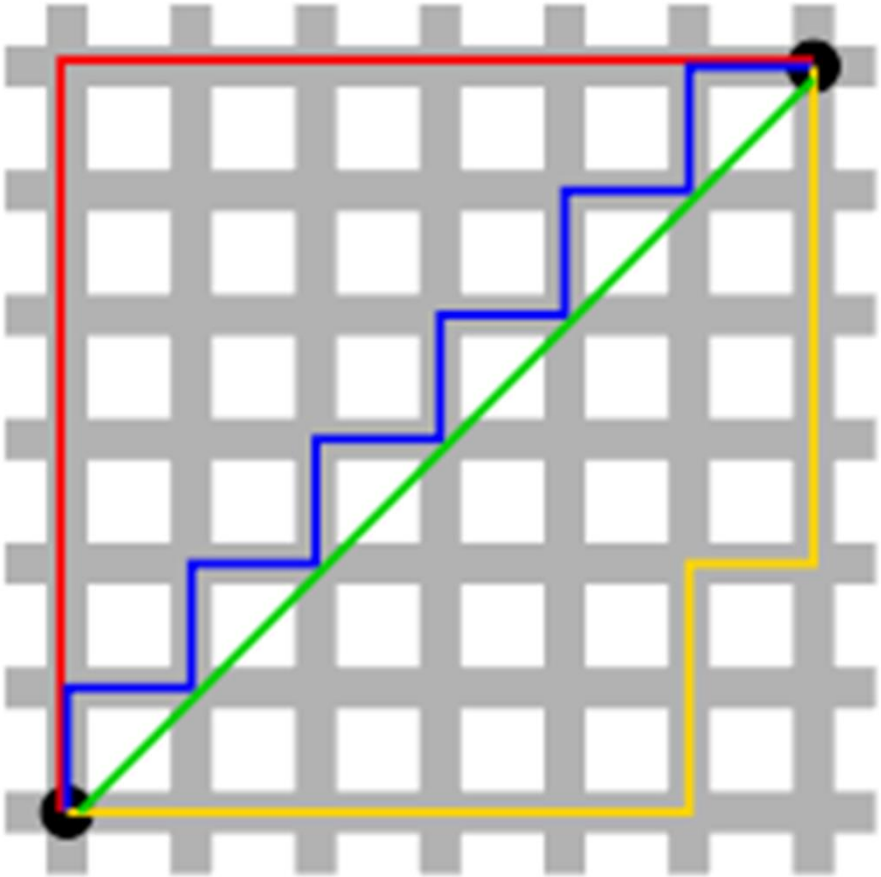
Distance Measurements

- Euclidian Distance



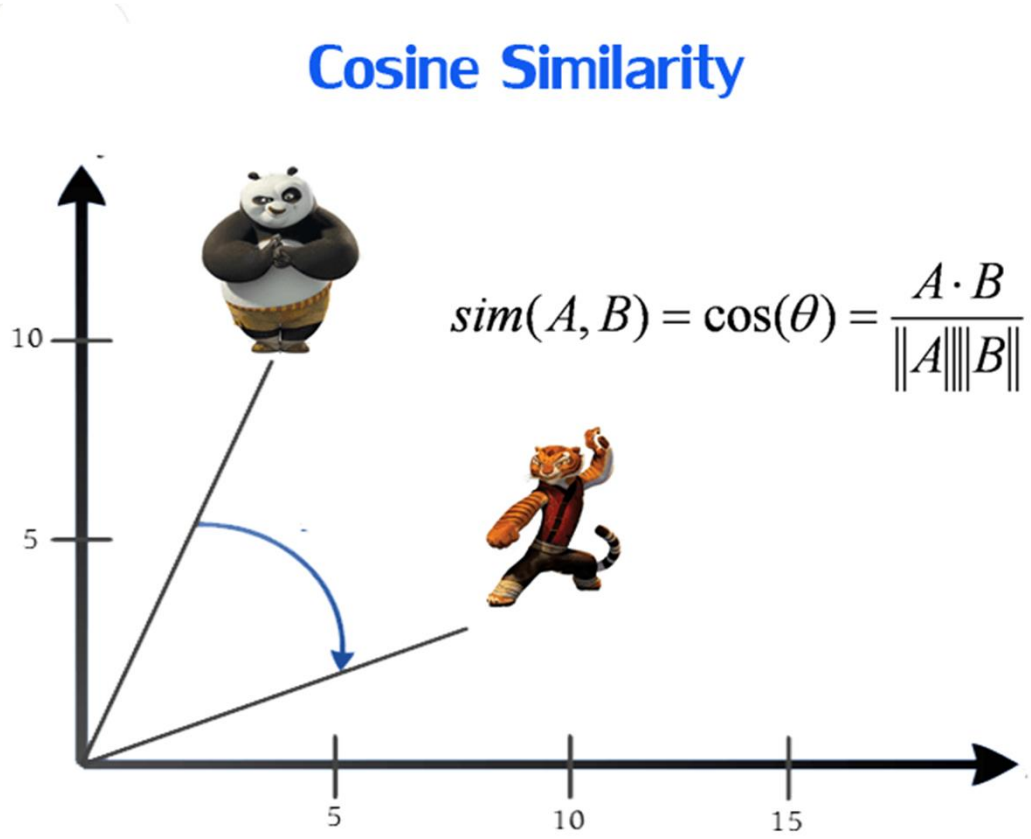
Distance Measurements

- Manhattan Distance



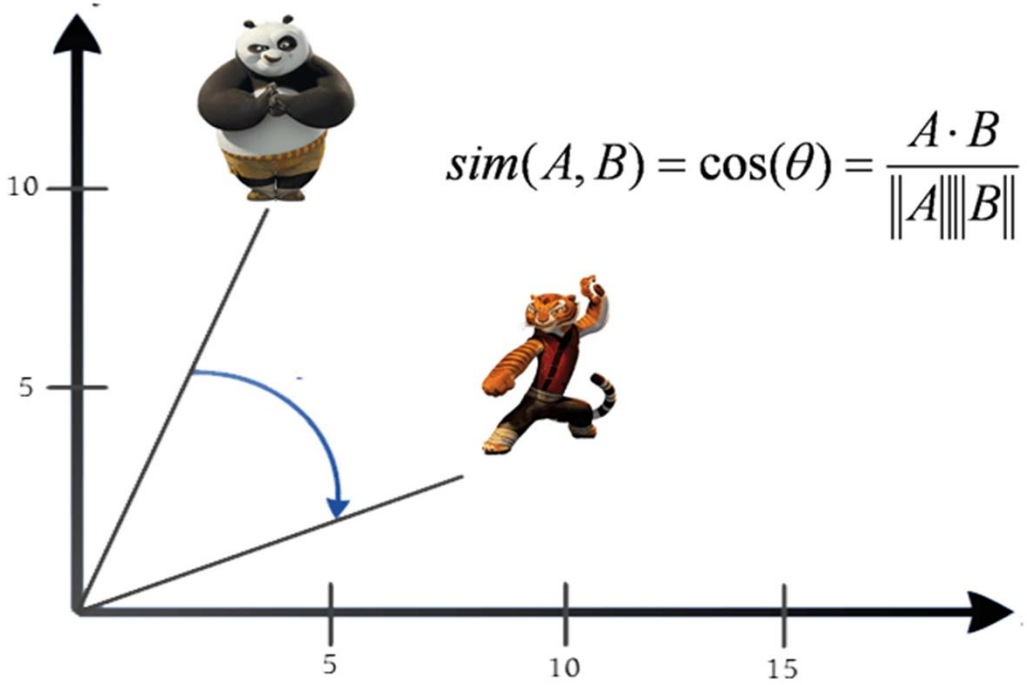
Distance Measurements

- Cosine Similarity



Distance Measurements

Cosine Similarity



$$sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Distance Measurements (Nominal Attribute)

- Percentage of similarity/dissimilarity

- Method 1: Simple matching
  - *m*: # of matches, *p*: total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- Method 2: Use a large number of binary attributes
  - Creating a new binary attribute for each of the *M* nominal states

Distance Measurements (Ordinal Attribute)

- Replace *an ordinal variable value* by its rank: $r_{if} \in \{1, \ldots, M_f\}$
- Map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

  - Example: freshman: 0; sophomore: 1/3; junior: 2/3; senior 1
    - Then distance: d(freshman, senior) = 1, d(junior, senior) = 1/3

Distance Measurements (Mixed Vector)

One may use a weighted formula to combine their effects:

$$d(i,j) = \frac{\sum_{f=1}^{p} w_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} w_{ij}^{(f)}}$$

- If f is numeric: Use the normalized distance

Resources

- Google
- https://www.coursera.org/learn/cluster-analysis/lecture/KnvRC/2-4-distance-between-categorical-attributes-ordinal-attributes-and-mixed-types

Any Questions