Nearest Neighbor Search

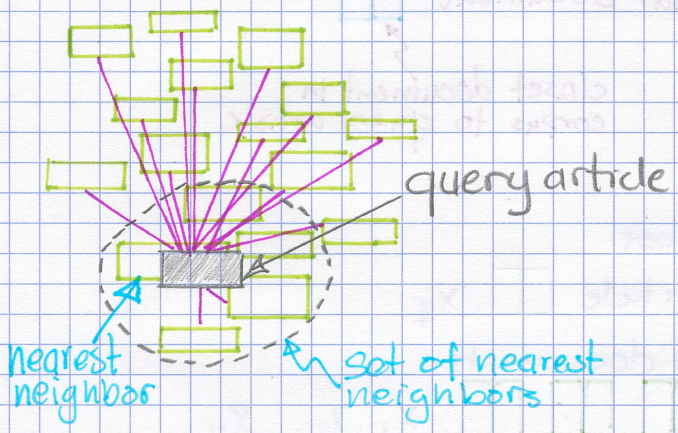
Retrieval as k-nearest neighbor search

Retrieving documents of interest
Document retrieval
- Currently reading article you like
- Goal: Want to find similar article

Challenges
- How do we measure similarity?
- How do we search over articles?

1-NN search for retrieval

Space of all articles, organized by similarity of text



1-NN algorithm

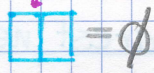
1 - Nearest neighbor
- Input: Query article x_q

Corpus of documents (N docs): $x_1, x_2, \ldots, x_N$

- Output: Most similar article $\leftarrow x^{NN}$

Formally: $x^{NN} = \min_{x_i} \text{distance}(x_q, x_i)$

1-NN algorithm

Initialize Dist2NN = ∞,  = $\emptyset$ (closest document)

For $i = 1, 2, \ldots, N$

Compute: δ = distance() (x_i, query document x_q)

If $\delta <$ Dist2NN (document i from corpus)

set  = x_i

set Dist2NN = δ

Return most similar document  x^{NN}

(closet document in corpus to query article)

k-NN algorithm

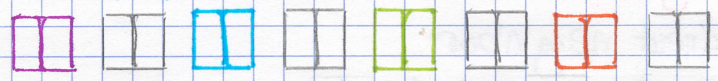
k - nearest neighbor

- Input: Query article  x_q

Corpus of documents

 $x_1, x_2, \ldots x_N$

- Output: List of k similar articles



Formally

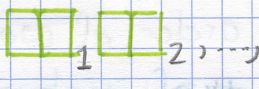
$x^{NN} = \{x^{NN_1}, \ldots, x^{NN_k}\}$

for all x_i not in x^{NN}, distance $(x_i, x_q) \geq$

$\geq \max_{x^{NN_j}, j=1,\ldots,k}$ distance (x^{NN_j}, x_q)

K-NN algorithm

Initialize Dist2kNN = sort($\delta_1, \ldots, \delta_k$) ← list of sorted distances (sort first k documents by distance to query doc)

= sort() ← list of sorted docs

For $i = k+1, \ldots, N$

Compute: δ = distance( , ) ← query doc

If δ < Dist2kNN[k] ← distance to k^{th} NN (furthest NN in set)

find j such that δ > Dist2kNN[j-1] but δ < Dist2kNN[j]

remove furthest article and shift queue:

 [j+1:k] =  [j:k-1]

Dist2kNN[j+1:k] = Dist2kNN[j:k-1]

(inserting new article)

set Dist2kNN = δ and  [j] = 

Return k most similar articles  ← closest k docs to query doc

Critical elements of NN search

Item (e.g., doc) representation

x_q ← 

Measure of distance between items:

δ = distance(x_i, x_q)