Lecture Note 7

March 16, 2018

1 topic

- K-shingles
- min-hash
- LSH

2 K-shingles

2.1 Intuition

Want to detect near duplicate web pages. But how do we measure similarity? To get the similarity of set A, B, we take Jaccard similarity: $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$. It is easy to see that we have $0 \leq J(A, B) \leq 1$.

2.2 Definition

K-shingle is a string of k consecutive characters

2.3 Example

- S_1: `"A boy jumped over the fence."`
- S_2: `"A fox jumped over the pool."`
- 3-shingles of S_1: $\{\texttt{A␣b}, \texttt{␣bo}, \texttt{boy}, \dots\}$

2.4 SHINGLING of documents

The choice of k need to be adapted according to possible number of documents. If $k = 1$, then there are only 27 different shingles, and it is very hard to distinguish documents. If $k = 5$, then we have $27^5 = 14M$ possible shingles. It may be enough for distinguishing emails, but may still be too small for longer documents. On the other hand, if k is very large, then we would need approximately $O(kn)$ space for n strings. This can be too much.

3 Minhash

3.1 Signature defined by

- random permutations: $\sigma_1, \sigma_2, \ldots, \sigma_n$
- $\sigma_i(S)$ = first element in σ_i that belongs to S
- signature of a set S: $sig(S) =< \sigma_1(S), \sigma_2(S), \ldots, \sigma_3(S) >$

3.2 Approximate Jaccard similarity

- $\text{Prob}(\sigma_1(A) = \sigma_1(B)) = J(A, B)$.

3.3 Algorithm

- For each row r
  - Compute $h_1(r), h_2(r), \ldots, h_k(r)$
  - For each column C
    - If c has 0 in row r, skip
    - If c has a 1 in row r, set $sig(i, c) \leftarrow \min(sig(i, c), h_i(r))$

3.4 Example

- Our universe is $X = \{Y_1, Y_2, \ldots, Y_7\}$
- $S_1 = \{Y_1, Y_3, Y_5\}$
- $S_2 = \{Y_1, Y_4, Y_5, Y_6\}$
- $\sigma_1 = Y_2, Y_1, Y_7, Y_4, Y_6, Y_3, Y_5$
- $\sigma_2 = Y_2, Y_3, Y_7, Y_4, Y_5, Y_6, Y_1$
- $\sigma_3 = Y_7, Y_5, Y_3, \ldots$
- $\sigma_4 = Y_6, Y_1, \ldots$
- $sig(S_1) =< Y_1, Y_3, Y_5, Y_1 >$
- $sig(S_2) =< Y_1, Y_3, Y_5, Y_6 >$

4 Distance Measure

- LP norm of vector
- Jaccard distance:
- Cosine distance: $\cos\theta = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|}$.
- Edit distance
- Hamming distance