IERG4300 Fall Tutorial 9

Hongyu DENG

Department of Information Engineering

The Chinese University of Hong Kong

Outline

- Minhash & LSH
- TP/TN FP/FN
- Q&A

Minhash & LSH

- Minhash Example
  - Consider we have 4 sets: S1, S2, S3 and S4.
    - S1 = {a, b, f, g}
    - S2 = {c, d, e}
    - S3 = {a, d, e, g}
    - S4 = {b, c, f}
  - π is a random policy.

Minhash & LSH

- Minhash Example

Feature	S1	S2	S3	S4
a	1	0	1	0
b	1	0	0	1
c	0	1	0	1
d	0	1	1	0
e	0	1	1	0
f	1	0	0	1
g	1	0	1	0

Change order

Feature	S1	S2	S3	S4
c	0	1	0	1
e	0	1	1	0
a	1	0	1	0
d	0	1	1	0
b	1	0	0	1
g	1	0	1	0
f	1	0	0	1

Minhash & LSH

- Minhash Example

Feature	S1	S2	S3	S4
c	0	1	0	1
e	0	1	1	0
a	1	0	1	0
d	0	1	1	0
b	1	0	0	1
g	1	0	1	0
f	1	0	0	1

(a) Policy 1

Feature	S1	S2	S3	S4
f	1	0	0	1
e	0	1	1	0
g	1	0	1	0
c	0	1	0	1
d	0	1	1	0
a	1	0	1	0
b	1	0	0	1

(b) Policy 2

Feature	S1	S2	S3	S4
g	1	0	1	0
f	1	0	0	1
c	0	1	0	1
d	0	1	1	0
b	1	0	0	1
a	1	0	1	0
e	0	1	1	0

(c) Policy 3

Minhash & LSH

- Minhash Example
  - We can estimate the similarity

Feature	S1	S2	S3	S4
a	1	0	1	0
b	1	0	0	1
c	0	1	0	1
d	0	1	1	0
e	0	1	1	0
f	1	0	0	1
g	1	0	1	0

Reduce Dim

π	S1	S2	S3	S4
Policy1	3	1	2	1
Policy2	1	2	2	1
Policy3	1	3	1	2

Minhash & LSH

- Prob($h(C_1) = h(C_2)$) = $sim(C_1, C_2)$

- Explanation
  - Suppose only two sets: S1 and S2
  - There are 3 types:
    - S1, S2 = 1 (A) with a rows
    - S1 = 1, S2 = 0 or S1 = 0, S2 = 1 (B) with b rows
    - S1, S2 = 0 (C) with c rows

Minhash & LSH

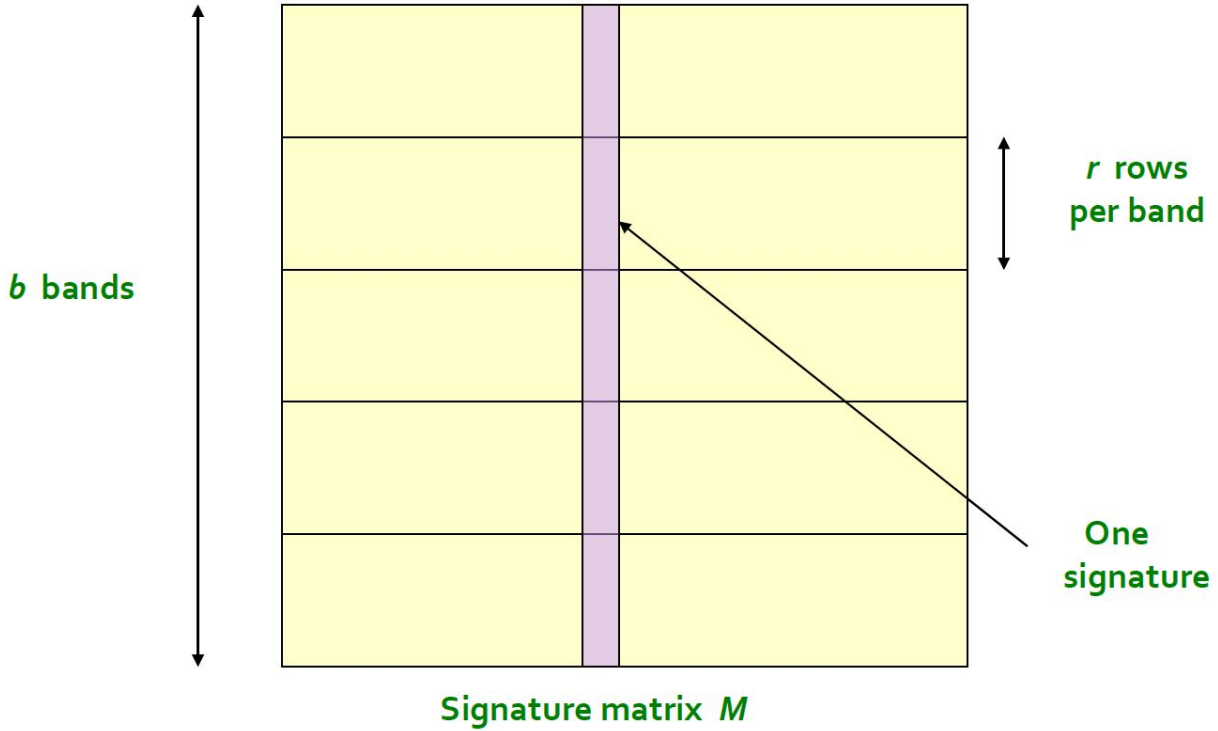
- Prob($h(C_1) = h(C_2)$) = $sim(C_1, C_2)$

- Explanation
  - sim(C1, C2) = a/(a + b +c)
  - Prob(h(C1) = h(C2)) means the probability of reaching a type-A row before a type-B or type-C row
  - Prob(h(C1) = h(C2)) = a/(a + b +c)

Minhash & LSH

- We want to find the relationship between r, b, similarity and probability in Q1(a).
    - First, we need to recall the **r** and **b** in a Matrix M.
        - To best understand, we give some specific value to them. Here we suppose r = 5, b = 20, s = 0.7.

Minhash & LSH



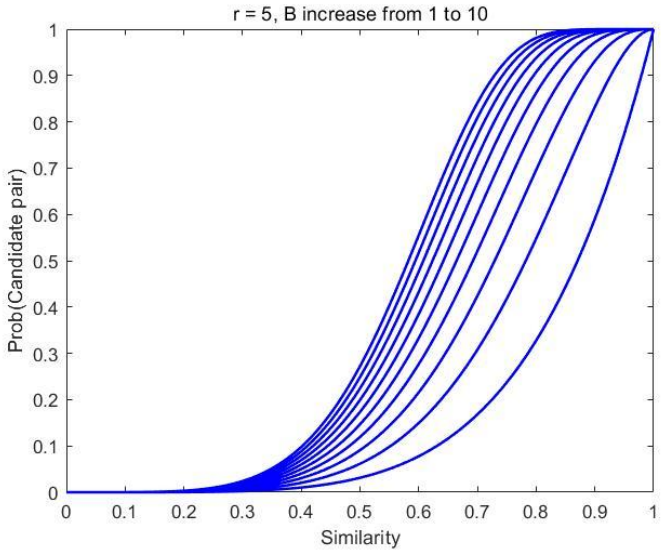
Screen Capture in Page48, Topic 4

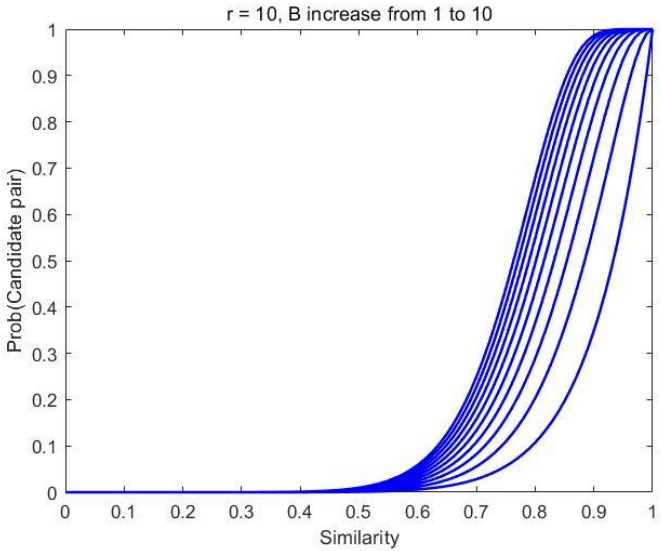
Minhash & LSH

- Second, we need to recall how to compute the similarity of two documents(C1 and C2).
  - In one band, we have 5 rows, so it is easy to obtain the probability C1, C2 identical in these bands.
    $(0.7)^5 = 0.1681$
  - Considering that we have 20 bands, the probability C1, C2 are not similar in whole bands.
    $(1 - (0.7)^5)^{20} = 0.0252$
  - Similarly, we can also obtain the probability C1, C2 identical in at least one of all bands.
    $1 - (1 - (0.7)^5)^{20} = 0.9748$

Minhash & LSH

- Then, we can plot the 2-D picture.
  - However, it is not a suitable method to draw a 2-D plot with similarity and probability.





Minhash & LSH

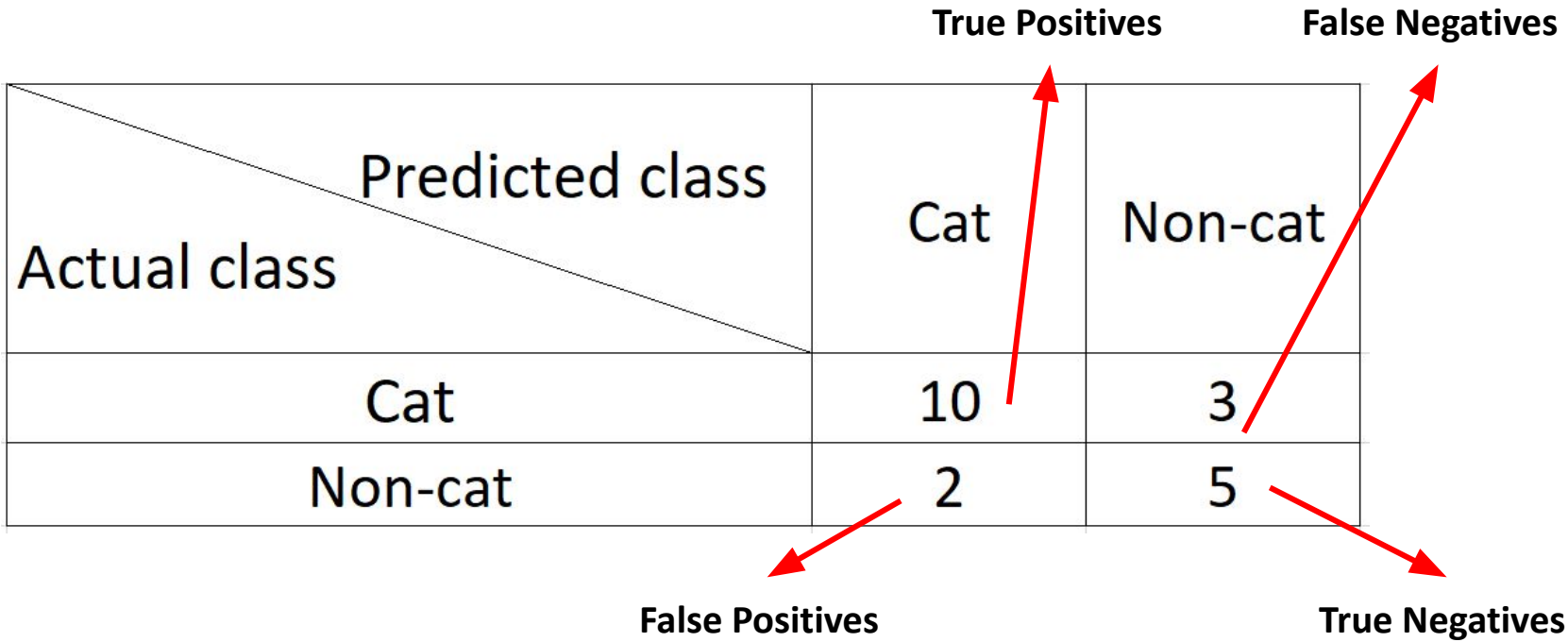
- Then, we can plot the 2-D picture.
  - You can try to find the relationship between **r** and **b**.
  - Try to draw the graph with **r** being the x-axis and **b** being the y-axis.

TP/TN FP/FN

- Confusion matrix
  - In confusion matrix, there are four part:
    - True Positives(TP) False Negatives(FN)
    - False Positives(FP) True Negatives(TN)

TP/TN FP/FN

- Confusion matrix



Q&A for Homework 3