IS622 Week 4 Homework

Ben Arancibia

September 19, 2015

3.1.3 Suppose we have a universal set U of n elements, and we choose two subsets S and T at random, each with m of the n elements. What is the expected value of the Jaccard similarity of S and T ?

Each item in T has an m / n chance of also being in S. The expected number of items common to S & T is therefore m^2 / n.

Exp. Jaccard Similarity = (No. of common items) / (Size of T + Size of S - Number of common items) = m / (2n - m) after simplification.

3.3.3

(a) Compute the minhash signature for each column if we use the following three hash functions: h1(x) = 2x + 1 mod 6; h2(x) = 3x + 2 mod 6; h3(x)=5x+2 mod6.

```
s1 <- c(0,0,1,0,0,1); s2 <- c(1,1,0,0,0,0); s3 <- c(0,0,0,1,1,0)
s4 <- c(1,0,1,0,1,0); element <- c(0,1,2,3,4,5)
h1 <- function(x) { (2*x + 1) %% 6 }
h2 <- function(x) { (3*x + 2) %% 6 }
h3 <- function(x) { (5*x + 2) %% 6 }

hashlist <- list(h1,h2,h3)
setlist <- list(s1,s2,s3,s4)

solution_3.3.3 <- computeMinhashSigs(hashlist, setlist)

#switch the binary to their corresponding values.

s1 <- c(2,5); s2 <- c(0,1); s3 <- c(3,4); s4 <- c(0,2,4)
setlist <- list(s1,s2,s3,s4)

solution_3.3.3 <- computeMinhashSigs(hashlist, setlist)
solution_3.3.3
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    5    1    1    1
## [2,]    2    2    2    2
## [3,]    0    1    4    0
```

It looks like the h2 has believes each row is identical while h1 believes they look nearly identical

(b) Which of these hash functions are true permutations?

```
hashlist <- list("h1"=h1,"h2"=h2,"h3"=h3)
row_count <- 5

hashPermuteDirect(hashlist, row_count)
```

```
##      [,1] [,2]
## [1,] "h1" "3"
## [2,] "h2" "2"
## [3,] "h3" "6"
```

(c) How close are the estimated Jaccard similarities for the six pairs of columns to the true Jaccard similarities?

Pretty close

3.5.5

Compute the cosines of the angles between each of the following pairs of vectors.

```r
angle <- function(x,y){
  dot.prod <- x%*%y
  norm.x <- norm(x,type="2")
  norm.y <- norm(y,type="2")
  theta <- acos(dot.prod / (norm.x * norm.y))
  as.numeric(theta)
}
```

(a) (3,−1,2)and(−2,3,1).

```r
x <- as.matrix(c(3,-1,2))
y <- as.matrix(c(-2,3,1))
angle(t(x),y)
```

```
## [1] 2.094395
```

(b) (1,2,3)and(2,4,6).

```r
x <- as.matrix(c(1,2,3))
y <- as.matrix(c(2,4,6))
angle(t(x),y)
```

```
## [1] 2.107342e-08
```

(c) (5,0,−4)and(−1,−6,2).

```r
x <- as.matrix(c(5,0,-4))
y <- as.matrix(c(-1,-6,2))
angle(t(x),y)
```

```
## [1] 1.893438
```

(d) (0,1,1,0,1,1) and (0,0,1,0,0,0).

```r
x <- as.matrix(c(0,1,1,0,1,1))
y <- as.matrix(c(0,0,1,0,0,0))
angle(t(x),y)
```

```
## [1] 1.047198
```