Curse of dimensionality

An illustration of problems with highly-dimensional data

Barbora Hladká and Martin Holub, Charles University in Prague – NPFL 054 (2017–18)

I. Distribution of random feature vector values

– each feature is binary, with the same Bernoulli distribution

```
N     <- 10^6           # number of observations
dim   <- 7              # number of dimensions
prob  <- 1/10           # probability of value 1

binom_vector <- character(N)
for(i in 1:N) binom_vector[i] <- paste(rbinom(dim,1,prob), collapse="")

expected_values = 2^dim
emerged_values = length(unique(binom_vector))

print( sort(table(binom_vector), dec=T) )
```

```
observations = 1,000,000
dimensions = 7
p(x = 1) = 0.1
number of possible different values = 128
number of emerged different values = 125

0000000 0010000 0100000 0001000 0000001 0000010 1000000 0000100 0001010 1001000
 477890   53566   53305   53303   53279   53272   53249   53023    6100    5995
1000100 0010001 0000011 0000101 0001001 0011000 0110000 0010100 0001100 0000110
   5981    5967    5961    5923    5901    5899    5898    5896    5893    5886
0100010 1000010 1100000 0101000 0100001 0100100 1010000 1000001 0010010 0010011
   5867    5844    5843    5815    5814    5810    5779    5778    5722     714
1001010 1000110 0010110 0101100 0100011 1001100 0001101 1100100 1010001 0001110
    704     700     692     690     686     680     673     673     672     669
0101010 1100010 1010010 0000111 0011100 0100110 0110100 0111000 0011010 0101001
    667     665     663     656     656     652     648     648     646     640
1001001 0110010 0010101 1010100 0011001 1100001 1000101 1101000 1110000 1000011
    639     637     635     635     634     634     632     630     630     626
0100101 0001011 1011000 0110001 1011010 1001011 0110110 1001101 1001110 1010101
    622     616     614     589      95      87      83      78      77      77
0010111 0101101 1011001 1011100 1101001 0011110 0111001 1101100 1110100 0011011
     76      76      75      75      75      73      73      73      73      72
0011101 1000111 1010011 1010110 1111000 0100111 0101011 1100011 1101010 1110001
     72      72      72      72      72      71      71      70      69      68
0110011 1100101 0110101 0101110 0111010 1110010 0001111 0111100 1100110 1011110
     66      66      65      63      62      62      56      56      55      14
1101011 1001111 1101110 0110111 0111101 0101111 1101101 0011111 0111011 1110101
     13      12      11      10      10       9       9       8       8       8
1110110 1010111 1111010 1110011 1111100 1011011 1100111 0111110 1011101 1110111
      8       7       7       6       6       5       5       4       4       3
1111001 0111111 1111011 1101111 1111110
      3       2       2       1       1
```

II. Randomly distributed points in a unit cube – distribution of distances from 0

```
dim   <- 6
n     <- 10000

cube <- data.frame(
        x1 = runif(n),
        x2 = runif(n),
        x3 = runif(n),
        x4 = runif(n),
        x5 = runif(n),
        x6 = runif(n)
      )

distances <- numeric(n)

for(i in 1:n) distances[i] <- sqrt(sum(cube[i,]^2))
greater_than_1 <- sum(distances > 1)

message("Most of the distances (",
        format(greater_than_1/n*100, digits=3), "%) are greater than 1.")

message("Frequency of distances in intervals:")
print(table(cut(distances, breaks=seq(0, 2.5, 0.5))))

----------------------------------------------------------------

This program generates 10,000 random 6-dimensional sample points in a unit cube.
Maximum possible distance from 0 is: 2.45

Distances from 0 in the sample of 10000 points:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.324   1.214   1.407   1.390   1.578   2.183

Most of the distances (91.7%) are greater than 1.
Frequency of distances in intervals:

(0,0.5] (0.5,1] (1,1.5] (1.5,2] (2,2.5]
      7     824    5591    3529      49
```

III. Randomly distributed points in a unit cube – distribution of mutual distances

```
dim   <- 6
n     <- 150
d     <- choose(n,2)
lim   <- 0.5

message("Maximum possible distance between two points is: ",
        format(sqrt(6), digits=3) )
message("Number of different pairs is: ", d)

cube <- data.frame(
        x1 = runif(n),
        x2 = runif(n),
        x3 = runif(n),
        x4 = runif(n),
        x5 = runif(n),
        x6 = runif(n)
      )

distances <- numeric(d)

k <- 1
for(i in 1:(n-1) ) for(j in (i+1):n ) {
   distances[k] <- sqrt( sum((cube[i,]-cube[j,])^2) ); k <- k+1
}
greater_than_lim <- sum(distances > lim)

message("Most of the distances (",
       format(greater_than_lim/d*100, digits=3), "%) are greater than ", lim, ".")

message("Frequency of distances in intervals:")
print(table(cut(distances, breaks=seq(0, 2.5, 0.25))))

----------------------------------------------------------------

This program generates 150 random 6-dimensional sample points in a unit cube.
Maximum possible distance between two points is: 2.45
Number of different pairs is: 11,175

Mutual distances in the sample of 150 points:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1173  0.8098  0.9797  0.9732  1.1420  1.8770

Most of the distances (97%) are greater than 0.5.
Frequency of distances in intervals:

 (0,0.25] (0.25,0.5] (0.5,0.75]   (0.75,1]   (1,1.25] (1.25,1.5] (1.5,1.75]
        9        322       1704       3914       3793       1301        128

  (1.75,2]   (2,2.25] (2.25,2.5]
         4          0          0
```