Curse of dimensionality

Mathias Bourel

18/3/2019

1 First Example

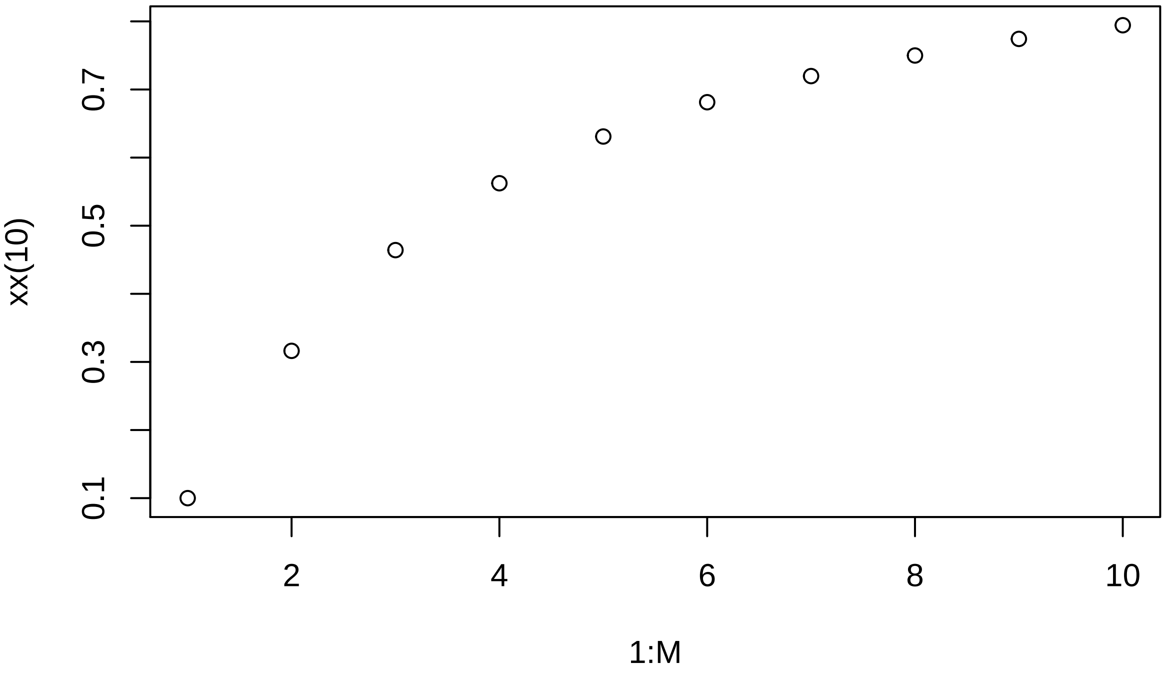
Suppose we send out a hypercubical neighborhood about a target point to capture a fraction r of the observations. Since this corresponds to a fraction r of the unit volume, the expected edge length will be $e_p(r) = r^{1/p}$. In ten dimensions $e_{10}(0.01) = 0.63$ and $e_{10}(0.1) = 0.80$, while the entire range for each input is only 1.0. So to capture 1% or 10% of the data to form a local average, we must cover 63% or 80% of the range of each input variable. Such neighborhoods are no longer "local". Reducing r dramatically does not help much either, since the fewer observations we average, the higher is the variance of our fit.

Capture 10% of the data:

```
M=10
xx=function(M){
  vec=NULL
  for(i in 1:M) { vec[i]=(0.1)^(1/i)}
return(vec)
}
xx(10)
```

```
##  [1] 0.1000000 0.3162278 0.4641589 0.5623413 0.6309573 0.6812921 0.7196857
##  [8] 0.7498942 0.7742637 0.7943282
```

```
plot(x=1:M,y=xx(10))
```



2- All of the Volume is in the corners

From https://mc-stan.org/users/documentation/case-studies/curse-dims.html

```
euclidean_length <- function(u) sqrt(sum(u * u));
euclidean_length(c(3, 4));
```

```
## [1] 5
```

Suppose we have a square and inscribe a circle in it, or that we have a cube and a sphere inscribed in it. When we extend this construction to higher dimensions, we get hyperspheres inscribed in hypercubes. This section illustrates the curious fact that as the dimensionality grows, most of the points in the hypercube lie outside the inscribed hypersphere.

Suppose we have an N-dimensional hypercube, with unit-length sides centered around the origin $0 = (0, \ldots, 0)$. The hypercube will have 2N corners at the points $\left(\pm\frac{1}{2}, \ldots, \pm\frac{1}{2}\right)$. Because its sides are length 1, it will have also have unit volume, because $1^N = 1$.

If $N = 1$, the hypercube is a line from $-\frac{1}{2}$ to $\frac{1}{2}$ of unit length (i.e., length 1). If $N = 2$, the hypercube is a square of unit area with opposite corners at $\left(-\frac{1}{2}, -\frac{1}{2}\right)$ and $\left(\frac{1}{2}, \frac{1}{2}\right)$. With $N = 3$, we have a cube of unit volume, with opposite corners at $\left(-\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}\right)$ and $\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)$, and unit volume. And so on up the dimensions.

Now consider the biggest hypersphere you can inscribe in the hypercube. It will be centered at the origin and have a radius of $\frac{1}{2}$ so that it extends to the sides of the hypercube. A point y is within this hypersphere if the distance to the origin is less than the radius, or in symbols, if $||y|| < \frac{1}{2}$. Topologically speaking, we have defined what is known as an open ball, i.e., the set of points within a hypersphere excluding the limit points at distance exactly $\frac{1}{2}$ (we could've worked with closed balls which include the limit points making up the surface of the ball because this surface (a hypersphere) has zero volume).

In one dimension, the hypersphere is just the line from $-\frac{1}{2}$ to $\frac{1}{2}$ and contains the entire hypercube. In two dimensions, the hypersphere is a circle of radius $\frac{1}{2}$ centered at the origin and extending to the center of all four sides. In three dimensions, it's a sphere that just fits in the cube, extending to the middle of all six sides.

Monte Carlo method's

We know the volume of the unit hypercube is one, but what is the volume of the ball in the inscribed hypersphere? You may have learned how to define an integral to calculate the answer for a ball of radius r in two dimensions as πr^2, and may even recall that in three dimensions it's $\frac{4}{3}\pi r^3$. But what about higher dimensions? We could evaluate harder and harder multiple integrals, but we'll instead use the need to solve these integrals as an opportunity to introduce the fundamental machinery of using sampling to calculate integrals. Such methods are called "Monte Carlo methods" because they involve random quantities and there is a famous casino in Monte Carlo. They are at the very heart of modern statistical computing (Bayesian and otherwise).

Monte Carlo integration allows us to calculate the value of a definite integral by averaging draws from a random number generator. (Technically, the random number generators we have on computers, like used in R, are pseudorandom number generators in the sense that they're underlyingingly deterministic; for the sake of argument, we assume they are random enough for our purposes in much the way we assume the functions we're dealing with are smooth enough).

To use Monte Carlo methods to calculate the volume within a hypersphere inscribed in a hypercube, we need only generate draws at random in the hypercube and count the number of draws that fall in the hypersphere—our answer is the the proportion of draws that fall in the hypersphere. That's because we deliberately constructed a hypercube of unit volume; in general, we need to multiply by the volume of the set over which we are generating uniform random draws.

As the number of draws increases, the estimated volume converges to the true volume. Because the draws are i.i.d., it follows from the central limit theorem that the error goes down at a rate of $O(1/\sqrt{n})$. That means each additional decimal place of accuracy requires multiplying the sample size by one hundred. We can get rough answers with Monte Carlo methods, but many decimal places of accuracy requires a prohibitive number of simulation draws.

We can draw the points at random from the hypercube by drawing each component independently according to $y_i \sim U\left(-\frac{1}{2}, \frac{1}{2}\right)$. Then we count the proportion of draws that lie within the hypersphere. Recall that a point y lies in the hypersphere if $||y|| < \frac{1}{2}$.

Example. We'll first look at the case where $N = 2$, just to make sure we get the right answer. We know the area inside the inscribed circle is πr^2, so with $r = 1/2$, that's $\frac{\pi}{4}$. Let's see if we get the right result.

```
N <- 2;
M <- 1e6;
y <- matrix(runif(M * N, -0.5, 0.5), M, N);
p <- sum(sqrt(y[ , 1]^2 + y[ , 2]^2) < 0.5) / M;
print(4 * p, digits=3);
```

```
## [1] 3.14
```

Now, let's generalize and calculate the volume of the hypersphere inscribed in the unit hypercube (which has unit volume by construction) for increasing dimensions.

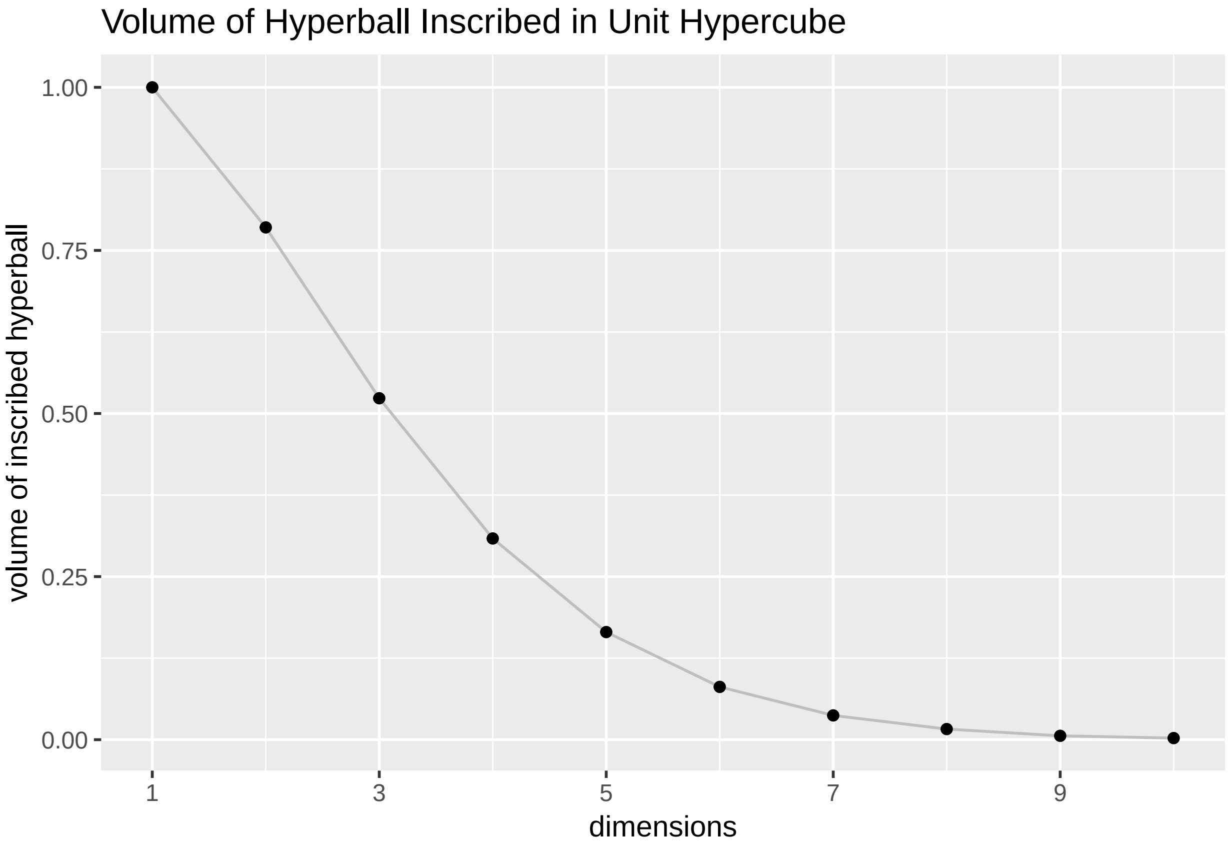
```
M <- 1e5;
N_MAX = 10;
Pr_inside <- rep(NA, N_MAX);
for (N in 1:N_MAX) {
  y <- matrix(runif(M * N, -0.5, 0.5), M, N);
  inside <- 0;
  for (m in 1:M) {
    if (euclidean_length(y[m,]) < 0.5) {
      inside <- inside + 1;
    }
  }
  Pr_inside[N] <- inside / M;
}
df = data.frame(dims = 1:N_MAX, volume = Pr_inside);
print(df, digits=1);
```

```
##    dims volume
## 1     1  1.000
## 2     2  0.785
## 3     3  0.523
## 4     4  0.308
## 5     5  0.165
## 6     6  0.081
## 7     7  0.037
## 8     8  0.016
## 9     9  0.006
## 10   10  0.002
```

Although we actually calculate the probability that a point drawn at random is inside the hyperphere inscribed in the unit hypercube, this quantity gives the volume inside the inscribed hypersphere.

Here's the result as a plot.

```r
library(ggplot2);
plot_corners <-
  ggplot(df, aes(x = dims, y = Pr_inside)) +
  scale_x_continuous(breaks=c(1, 3, 5, 7, 9)) +
  geom_line(colour="gray") +
  geom_point() +
  ylab("volume of inscribed hyperball") +
  xlab("dimensions") +
  ggtitle("Volume of Hyperball Inscribed in Unit Hypercube")
plot_corners;
```


Volume of Hyperball Inscribed in Unit Hypercube
volume of inscribed hyperball
dimensions