Information Retrieval in High Dimensional Data: Assignment 1

Benjamin Braun (03680075), Juri Fedjaev (03628226),
Nirnai Rao (03692571), Alexandros Sivris (03627456)

18.05.2017

Contents

Curse of Dimensionality

Task 1

See PYTHON-file in attachment.
The minimal angle can range between 0°and 180°. From the plot we see that up to approximately 50 dimensions the angle variance is not as large, because the average angle for the maximum variance would be about 90°(which is the mean angle of the full range, i.e. 0°- 180°). As the dimensionality increases the average minimum angle converges to a number very close to 90°and so a larger spectrum of angles is covered. As the plots in figure 0.1 indicate, the number of samples makes little to no difference in this behavior. This shows one aspect of the curse of dimensionality, because in higher dimensions all neighboring vectors are 90°apart. Therefore, a k-nearest neighbors algorithm would provide no useful results.

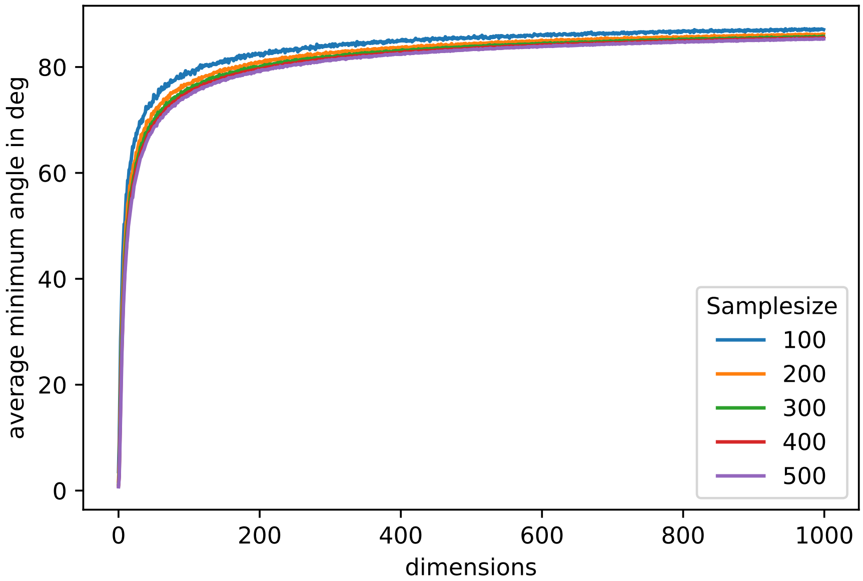


Figure 0.1: Average minimum angle against dimension d for different sample sizes.

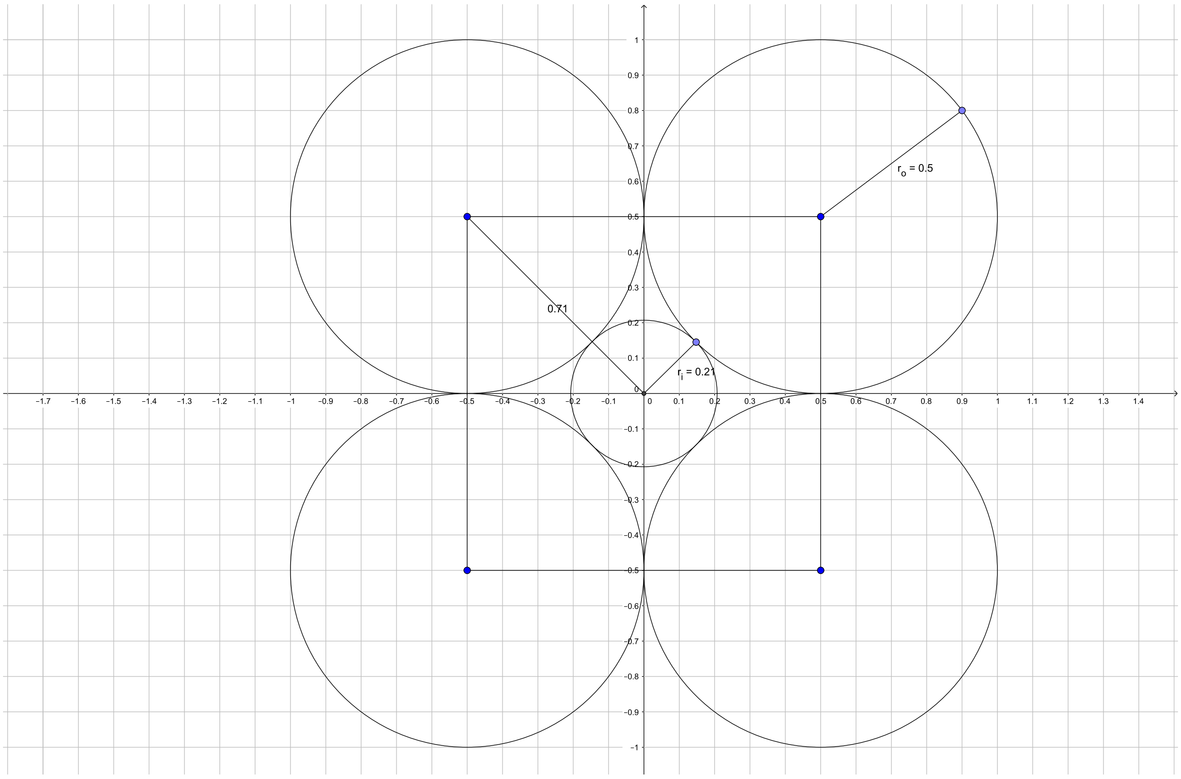


Figure 0.2: four circles with radius $r_o = \frac{1}{2}$ placed at the corners of a square of side length $a = 1$

Task 2

- From the figure below it is evident, that the sum of the radii of inner and outer circle equals half of the diagonal d of the square.

$$r_i = \frac{d}{2} - r_o = \frac{\sqrt{2} \cdot a}{2} - r_0 \tag{0.1}$$

  With equation 0.1, $r_i = 0.21$.

- In higher dimensions D the equation changes to the following:

$$r_i = \frac{d}{2} - r_o = \frac{\sqrt{D} \cdot a}{2} - r_0 \tag{0.2}$$

  This shows that the inner sphere grows with growing dimensions. The rate is proportional to $\sqrt{D}$.

- Case $D = 4$, $r_0 = frac12$:

$$r_i = \frac{\sqrt{4}}{2} - \frac{1}{2} = \frac{1}{2} \tag{0.3}$$

- Case $D = 9$, $r_0 = frac12$:

$$r_i = \frac{\sqrt{9}}{2} - \frac{1}{2} = 1 \tag{0.4}$$

- Case $D = 100$, $r_0 = frac12$:

$$r_i = \frac{\sqrt{100}}{2} - \frac{1}{2} = 4.5 \tag{0.5}$$

  For growing D, the radius r_i grows as well, but very slowly compared to the dimension due to the square root (cf. figure 0.3)

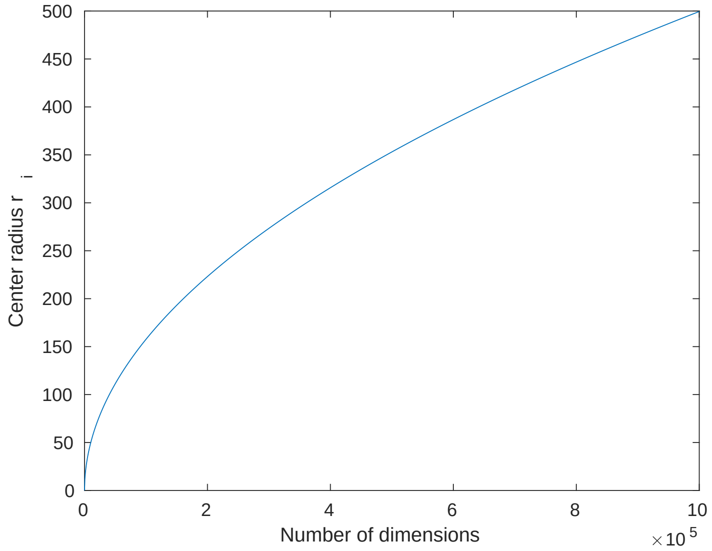


Figure 0.3: Plot of center radius r_i with growing dimensionality.

Statistical Decision Making

Task 3

- In figure 0.1 the sum of the probabilities of the random variables X and Y have been calculated. Since both equal one, it can be said that this is a true probability distribution.

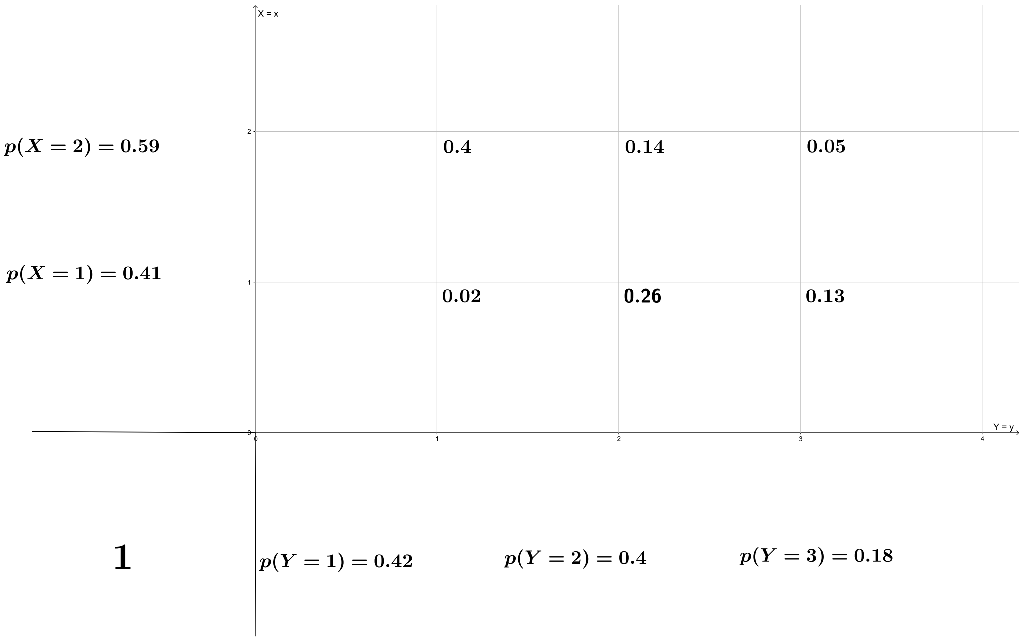


Figure 0.1: Probability Distribution and Sum of Probabilities equals 1

- $$E_{Y|X=2}[Y] = \frac{\sum_{y \in Y} y \cdot p(Y|X=2)}{P(X=2)} = \frac{0.4 \cdot 1 + 0.14 \cdot 2 + 0.05 \cdot 3}{0.59} = 1.41 \tag{0.1}$$

$$P(X = 1|Y = 3) = \frac{P(X = 1; Y = 3)}{P(Y = 3)} = \frac{0.13}{0.18} = 0.72 \tag{0.2}$$

- For $p(x; y)$ to be a joint probability density function, it needs to fulfill following criteria:

$$\iint p(x; y)\, dx\, dy \overset{!}{=} 1 \tag{0.3}$$

For the given joint density function the integral looks as follows:

$$\int_0^{\frac{1}{2}} \int_0^1 1 \, dx \, dy = \frac{1}{2} \neq 1 \tag{0.4}$$

The result from equation 0.4 shows that the given joint density function can not be a probability distribution.

- The marginal density functions from a joint density function are defined as follows:

$$p(x) = \int_Y p(x;y) \, dy \tag{0.5}$$

$$p(y) = \int_X p(x;y) \, dx \tag{0.6}$$

For given $p(x;y)$ the marginals calculate to:

$$p(x) = \int_0^\infty 2e^{-(x+y)} \, dy = \left[-2e^{-(x+y)}\right]_0^\infty = 2e^{-x} - \lim_{y\to\infty} 2e^{-(x+y)} = 2e^{-x} \tag{0.7}$$

$$p(y) = \int_0^y 2e^{-(x+y)} \, dx = \left[-2e^{-(x+y)}\right]_0^y = 2(e^{-y} - e^{-2y}) \tag{0.8}$$

- With $p_Y(y) = \frac{1}{15}(9+12y)$ and $p_Y(y = \frac{1}{2}) = 1$:

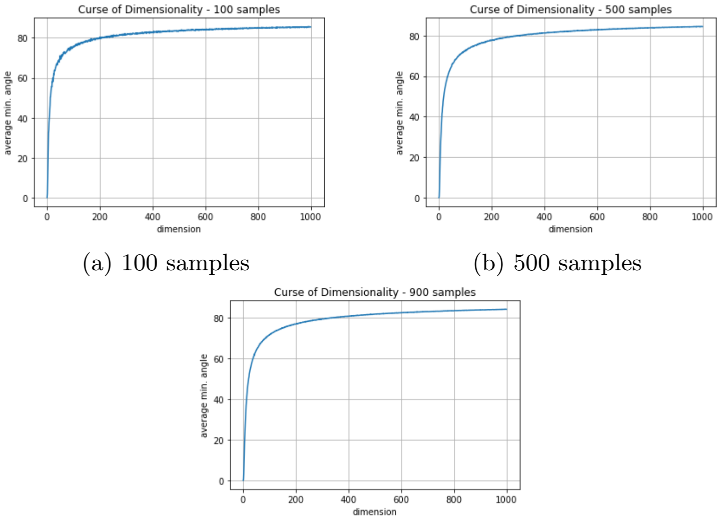
$$P\left(X \leq 2|Y = \frac{1}{2}\right) = \int_0^2 \frac{p_{X,Y}(x, y = \frac{1}{2})}{p_Y(y = \frac{1}{2})} dx \tag{0.9}$$

$$= \int_0^2 \frac{1}{15}\left(2x + 4 \cdot \frac{1}{2}\right) dx \tag{0.10}$$

$$= \frac{1}{15}\left[x^2 + 2x + C\right]_0^2 \tag{0.11}$$

$$= \frac{8}{15} \tag{0.12}$$

Appendix



(a) 100 samples

(b) 500 samples

(c) 900 samples

Figure 0.1: Average minimum angle for up to 1000 dimensions.