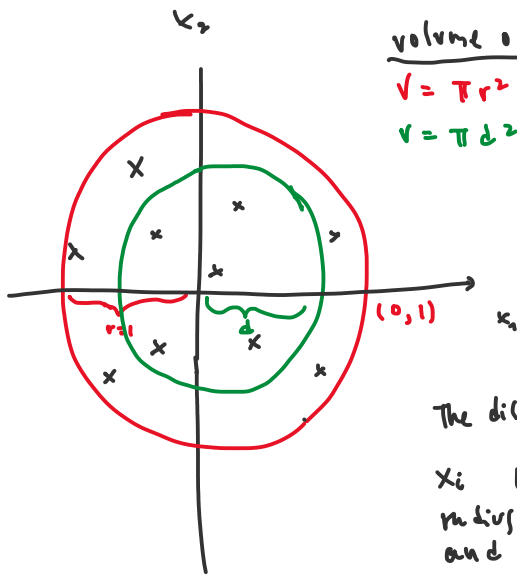


each a p -length vector

Suppose n observations $x_1, \dots, x_n$ are uniformly distributed on a p -dimensional unit ball, centered at origin. What is the median distance from the origin to the closest point?

For $p=2$:



volume of disk:

$$V = \pi r^2 = \pi$$

$$V = \pi d^2$$

The distance of x_i to the origin: $y_i = \sqrt{x_{i1}^2 + x_{i2}^2}$

x_i lies in the circle with radius $r=1$ with probability 1. and in the circle with radius d with probability $\pi d^2 / \pi$

Let y_i be the distance from x_i to the origin

$$y_i = \|x_i\|_2 = \left(\sum_{j=1}^p x_{ij}^2 \right)^{\frac{1}{2}}$$

Let $m_n = \min(y_1, \dots, y_n)$ be the closest point.

We want to find the median distance m of the closest point, defined by

$$P(m_n \leq \underline{m}) = \frac{1}{2}$$

(*)

Find an expression for the CDF of m_n :

$$\begin{aligned}
 P(m_n \leq m) &= 1 - P(m_n > m) && \text{all } y_i \text{'s larger than } m \\
 &= 1 - P(y_1 > m \cap y_2 > m \cap \dots \cap y_n > m) \\
 &\stackrel{\text{iid}}{=} 1 - P(y_i > m)^N \\
 &= 1 - [1 - P(y_i \leq m)]^N
 \end{aligned}$$

random variable fixed constant

Find an expression for the CDF of y_i :

$$\begin{aligned}
 P(y_i \leq y) &= P(x_{i1} \leq y \cap x_{i2} \leq y \cap \dots \cap x_{ip} \leq y) \\
 &= [P(x_{ij} \leq y)]^p \\
 &= y^p
 \end{aligned}$$

CDF of a univariate uniform variable
 $P(U \leq u) = u$

b) or the ratio between the volume of a ball with radius y and radius 1:

$$\frac{\frac{\pi^{\frac{p}{2}} y^p}{\Gamma(\frac{p}{2} + 1)}}{\frac{\pi^{\frac{p}{2}} 1^p}{\Gamma(\frac{p}{2} + 1)}} = y^p$$

Now, we can solve for the median distance in (*):

$$\begin{aligned}\frac{1}{2} &= P(m_N \leq m) \\ &= [1 - P(Y_i \leq m)]^N \\ &= [1 - m^p]^N\end{aligned}$$

$$\begin{aligned}\Rightarrow m^p &= 1 - \left(\frac{1}{2}\right)^{\frac{1}{N}} \\ m &= \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{N}}\right)^{\frac{1}{p}}\end{aligned}$$

This is equal to ERL eq. 2.24

ESL 2.7 a

$$\begin{aligned}
 x_i &\sim h(x) \\
 \epsilon_i &\sim N(0, \sigma^2) \\
 y_i &= f(x_i) + \epsilon_i
 \end{aligned}$$

The estimator $\hat{f}$ is linear in y_i :

$$\hat{f}(x_0) = \sum_{i=1}^N l_i(x_0; \mathcal{X}) y_i$$

$\uparrow$
 Ex: $\{l_i\}_{i=1}^N$ set of binary sample predictors

(a) Show that (i) linear regression (ii) KNN are members of this class of estimators, and describe the weights $l(x_0; \mathcal{X})$.

(ii) KNN

Let $H_K(x_0; \mathcal{X})$ denote the set with the K closest points to x_0 in $\mathcal{X}$.

$$\begin{aligned}
 \hat{f}(x_0) &= \frac{1}{K} \sum_{i=1}^N \mathbb{1}(x_i \in H_K(x_0; \mathcal{X})) y_i \\
 &= \sum_{i=1}^N l_i(x_0; \mathcal{X}) y_i
 \end{aligned}$$

indicator-function

where

$$l_i(x_0; \mathcal{X}) = \begin{cases} \frac{1}{K} & \text{if } x_i \in H_K(x_0; \mathcal{X}) \\ 0 & \text{otherwise} \end{cases}$$

(i) Linear Regression

Let X be the $N \times p$ design matrix of training sample and y the $N \times 1$ vector with training sample outcomes. Then

$$\begin{aligned}\hat{f}(x_0) &= x_0^T \hat{\beta} \\ &= \underbrace{x_0^T (X^T X)^{-1} X^T}_{1 \times N} y \\ &= \sum_{i=1}^N l_i(x_0; x) y_i\end{aligned}$$

where $l_i(x_0; x)$ is the i th element in the $1 \times N$ vector $x_0^T (X^T X)^{-1} X^T$.

Example:

Consider the simplest of simple cases, with only one regressor $x = x_1$ and no intercept:

$$\begin{aligned}\underbrace{x_0}_{1 \times 1} \underbrace{(X^T X)^{-1}}_{\sum_{i=1}^N x_i^2} \underbrace{X^T}_{1 \times N} &= \left\{ x_0 (\sum_{i=1}^N x_i^2)^{-1} x_i \right\}_i \\ &= \left\{ l_i(x_0; x) \right\}_i\end{aligned}$$

$$\hat{f}(x_0) = x_0 \sum_{i=1}^N l_i(x_0; x) y_i = x_0 \underbrace{\sum_{i=1}^N \frac{x_i y_i}{\sum x_i^2}}_{\hat{\beta}}$$