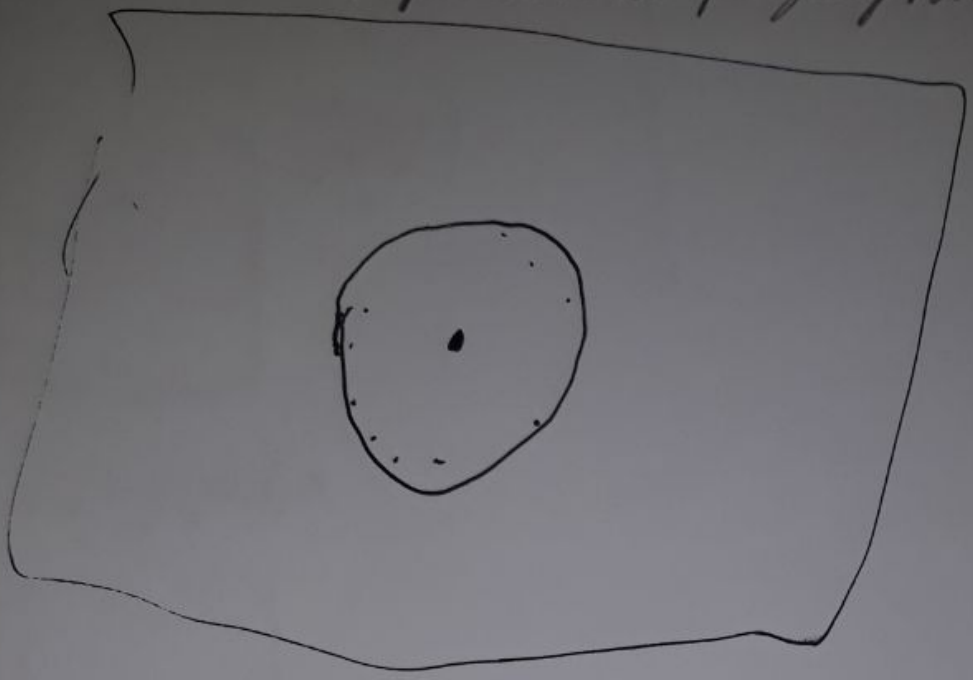
KNN

\+ 1). Много ключей, мало представителей
2) Не обязательно выделять признаки
3) Надстройка над более сложными моделями

— 1) Нужна вся выборка
2) Проклятье размерности
3) Долго работает

± Чувствителен к масштабу

Проклятие размерности.

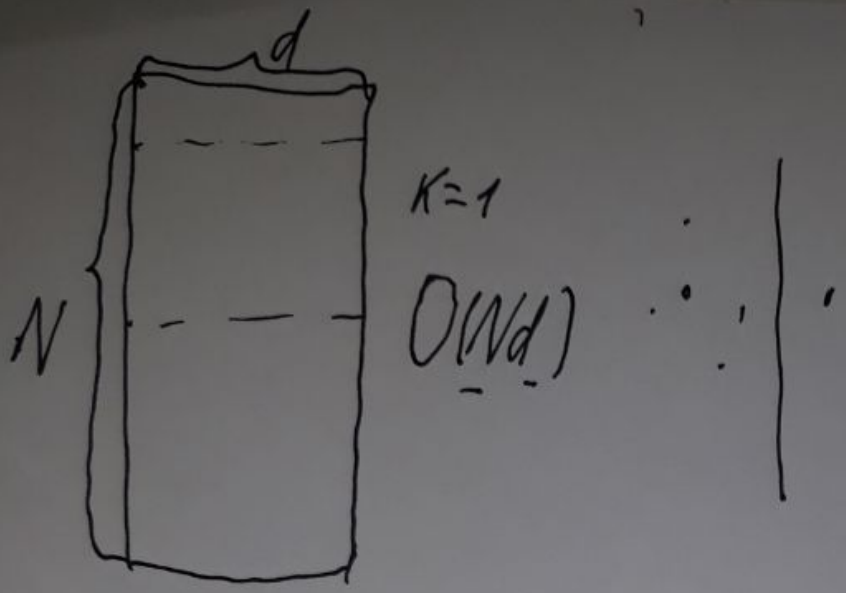


$V^n_{шара} = \pi R^2, \frac{4}{3}\pi R^3$

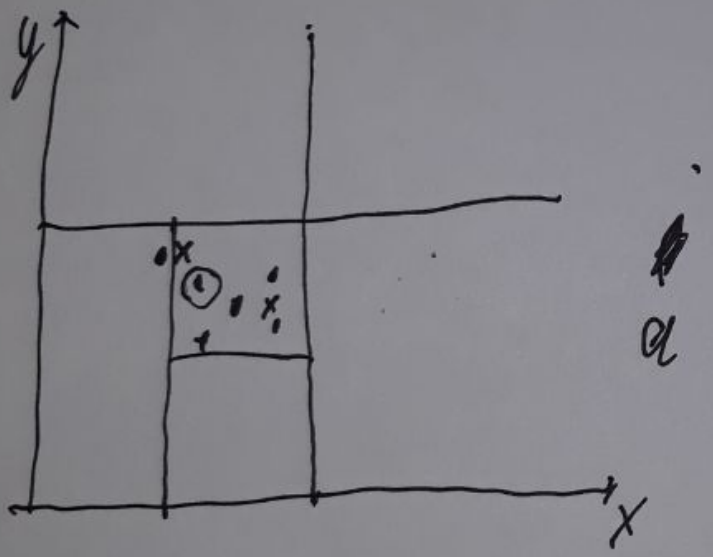
$C R^n$

$R-\varepsilon, \varepsilon > 0.$

$$\frac{C_1 R-\varepsilon^n}{C_2 R^n} = C_3 \left(\frac{R-\varepsilon}{R}\right)^n \xrightarrow[n\to\infty]{} 0$$



KD-деревья



d=10-20 - логарифмическое время

d>>20 - долго

Locality sensitive hashing (LSH)

~~$F(x_1, x_2)$~~

$x_1, x_2 \in X$

$F(x_1) = F(x_2)$

x_1 и x_2 близко

$F(x_1) \neq F(x_2)$ далеко.

Опр: семейство F называется (d_1, d_2, p_1, p_2) - чувствительным если $\forall x_1, x_2 \in X$:

если $\rho(x_1, x_2) \leq d_1$, то

$P(f(x_1) = f(x_2)) \geq p_1$.

если $\rho(x_1, x_2) \geq d_2$ то

$P(f(x_1) = f(x_2)) < p_2$

$p_1 \geq p_2$ $d_1 \leq d_2$.

Метрика Жаккарда
(Jaccard)

A, B – множества

$$\rho_J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

$U = \{u_1, u_2, \ldots, u_n\}$.

$A, B \subset U$

$\pi = \{u_{k1}, u_{k2}, u_{k3}, u_{kn}\}$

$f_\pi(A) = \min\{\pi(u_i) \mid u_i \in A\}$.

$F = \{f_\pi \mid \pi \in \text{ПЕРЕСТ}(U)\}$.

Утв: вероятность что $\forall f_\pi \in F$ будет одинаковой на A, B равно $\frac{|A \cap B|}{|A \cup B|} = 1 - \rho_J(A,B)$

1) $u \in A, u \in B$ – p объектов

2) $u \in A, u \notin B$ или $u \notin A, u \in B$. – q объектов

3) $u \notin A, u \notin B$

$$\frac{p}{p+q} = 1 - \rho_J(A,B)$$

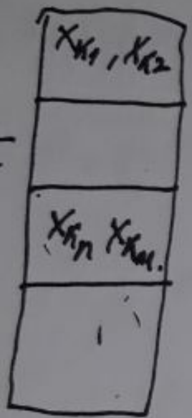
$\rho_J(A,B) \leq d_1$.

$1-\rho_J(A,B) \geq 1-d_1$.

$p_1 > 1-d_1$. $p_2 < 1-d_2$.

$(d_1, d_2, 1-d_1, 1-d_2)$ - чувствит.

$f \in F$



$K_{nearest} = 5$

объекта и что делать?

$g(x) = (f_1(x), f_2(x), \ldots, f_m(x))$.

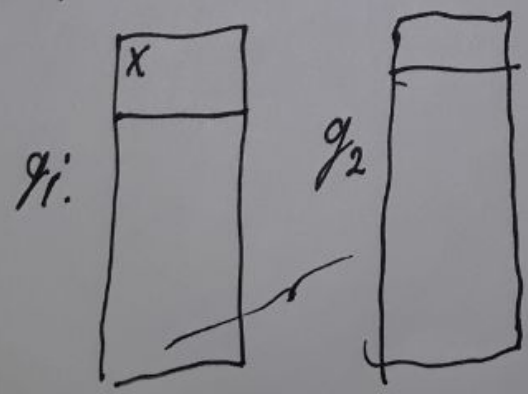
$f_1, \ldots, f_m \in F$

$g_1(x)$

$g_2(x)$

$\vdots$

$g_L(x)$



m, L —

Вер-ть того что мухи все испортит

$x_1 = 0.1 \quad y_1 = 1$

$x_2 = 0.5 \quad y_2 = 2$

$U = (0,0)$

второй угол $\sim U[0,1]$

$$P(\rho(U,x_1) > \rho(U,x_2)) =$$

$$P(\sqrt{0.1^2 + \xi_1^2} > \sqrt{0.5^2 + \xi_2^2}) =$$

$$= P(0.1^2 + \xi_1^2 > 0.5^2 + \xi_2^2) =$$

$$= P(\xi_1^2 \geq 0.24 + \xi_2^2) =$$

$$= \int_0^1 \int_{\sqrt{x_2^2 + 0.24}}^1 dx_1\, dx_2 \approx 0.275$$