L7: Approximate Nearest Neighbors

Jeff M. Phillips

January 29, 2020

Euclidean Distance $d(a,b) = \|a-b\|_2$

$a, b \in \mathbb{R}^d \qquad a = (a_1, a_2, \dots a_d)$

Raw Data $\longrightarrow \mathbb{R}^d$ usually d large

Example $\rightarrow$ bag-of-word

$\rightarrow$ word-vector embedding

$\rightarrow$ images

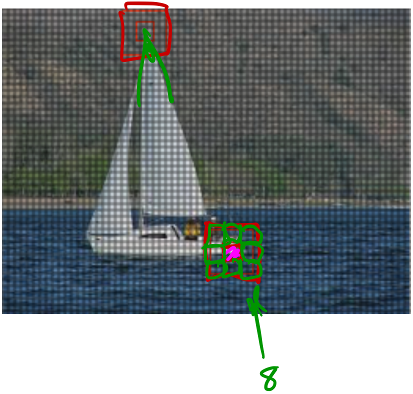
CNN $\longrightarrow$ intermediate layer

SIFT vector $\in \mathbb{R}^{128}$

Images and SIFT Features

N1	N2	N3
N8	X	N4
N7	N6	N5

Images and SIFT Features



N1	N2	N3
N8	X	N4
N7	N6	N5

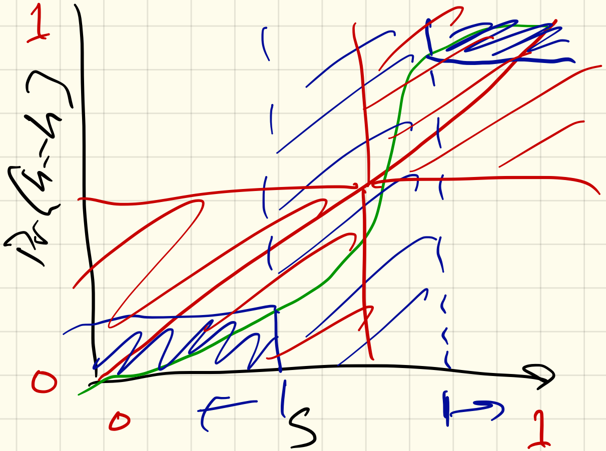
LSH for Euclidean dist.

$$\Pr_{h \in \mathcal{H}}\left[h(a) = h(b)\right] = S(a,b)$$

Jaccard

triangle

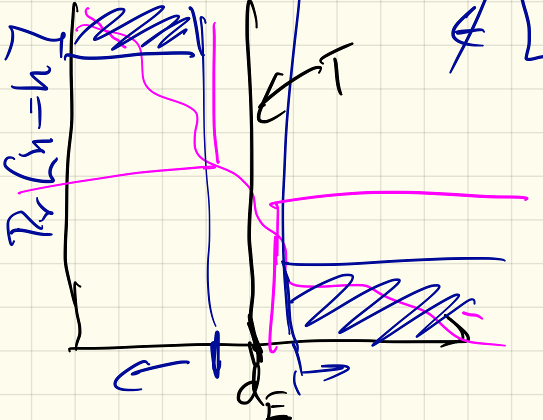
angular



Euclidean

$$S(a,b) = \langle a,b \rangle = \sum_{i=1}^{d} a_i b_i$$

$$\notin [0,1]$$



$h_{u,\eta}(x)$ $\qquad x \in \mathbb{R}^d$

$u \sim \mathrm{Unif}(\mathbb{S}^{d-1})$

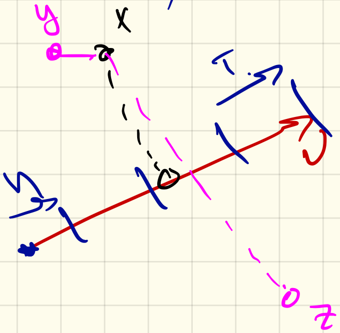
$\uparrow$ unit vectors in $\mathbb{R}^d$

$\|u\| = 1$

$\eta \in \mathrm{Unif}(0, T)$

$\uparrow$ dist threshold

$$h_{u,\eta}(x) = \left\lfloor \left(\frac{(\langle x, u\rangle + \eta)}{T} \right) \pmod{m} \right\rfloor$$



Uniform $\underline{u} \in \text{Unif}(S^{d-1})$

Gaussian RV.

$g \in \frac{1}{(2\pi)^{d/2}} e^{-\frac{\|x\|^2}{2}}$

$u \in \frac{g}{\|g\|}$

$g \in G_d$

$g = (g_1, g_2, \dots g_d)$

each $g_i \in G_1$

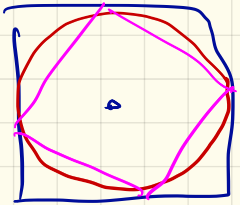
Box-Muller

$g = \sqrt{-2\ln(u_1)}\cos(2\pi u_2)$

$u \in \mathbb{R}^d$

$$\tilde{u} = \begin{bmatrix} \tilde{u}_1 \in \text{Unif}(0,1) \\ \tilde{u}_2 \in \text{Unif}(0,1) \\ \vdots \\ \tilde{u}_d \in \text{Unif}(0,1) \end{bmatrix}$$

$u \in \frac{\tilde{u}}{\|\tilde{u}\|}$



vol ball radius 1 in $\mathbb{R}^d$

box rad 1

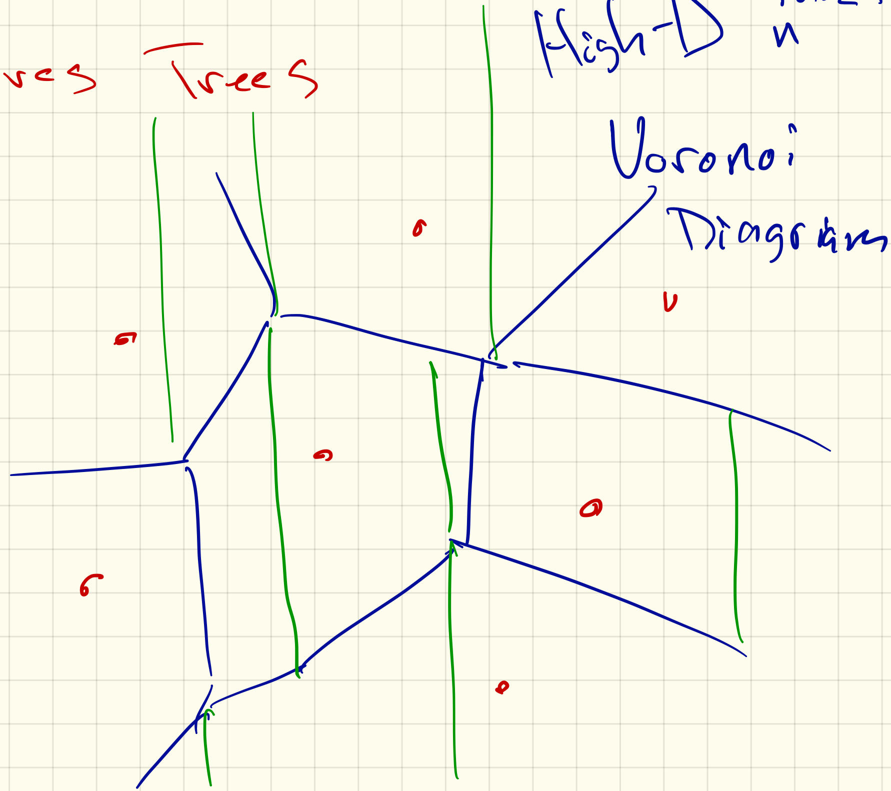
2 . . . 2 box

$1 = 2^d$

$$\frac{\pi^{d/2}}{\Gamma(d/2+1)} \approx \frac{\pi^{d/2}}{(d/2)!} \leq 1$$

High-d NN Euclidean

- LSH
- Data Structures Trees
- Graphs

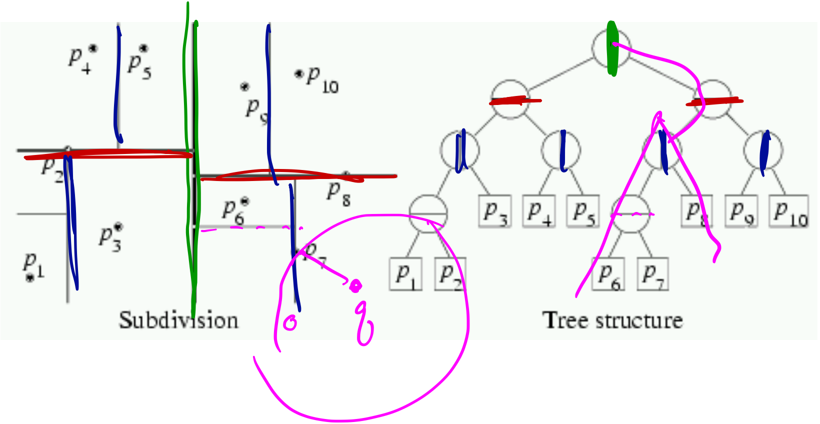


High-D $n^{\lceil d/2 \rceil}$

Voronoi Diagram

*k*D-Tree

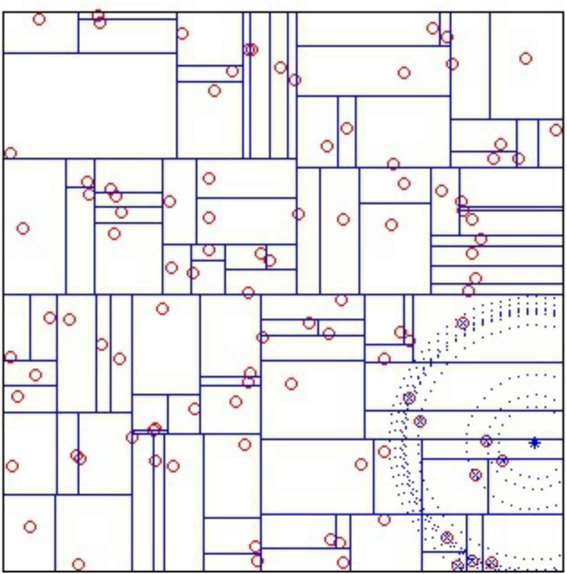
maybe d=12



Subdivision

Tree structure

Approximate Queries on *k*D-Tree

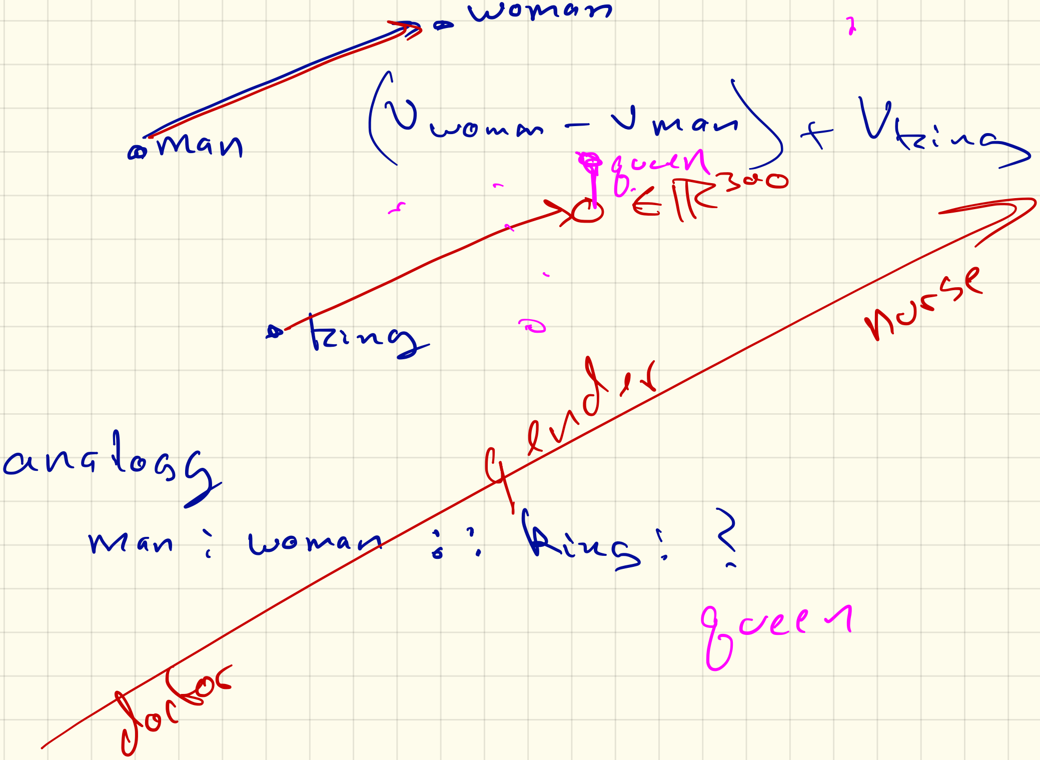


Word Vectors & Association

word = nurse → $v_{nurse} \in \mathbb{R}^{300}$

doctor $v_{doctor} \in \mathbb{R}^{300}$

if text context is similar the two words should have small cosine distance



$(v_{woman} - v_{man}) + v_{king}$

queen $\in \mathbb{R}^{300}$

analogy

man : woman :: king : ?

queen