L19: Noise in Data

- Random Projections
- Outliers
- Matrix Completion

Random Projections

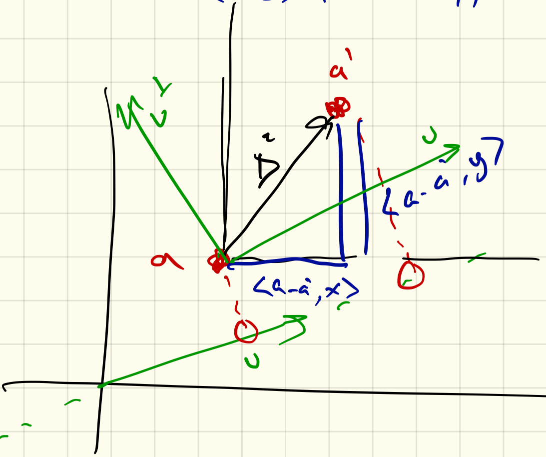
Input: Data $A \in \mathbb{R}^{n \times d} = \{a_1, a_2 \dots a_n\} \subset \mathbb{R}^d$

Goal: map $\mu : \mathbb{R}^d \to \mathbb{R}^k$ $(k \ll d)$

preserve all distances say $a, a' \in A$

$$(1-\varepsilon)\|a - a'\| \le \|\mu(a) - \mu(a')\| \le (1+\varepsilon)\|a - a'\|$$

w.p. $1-\delta$ ← prob. failure



$$\Delta^2 = \langle a-a', x\rangle^2 + \langle a-a', y\rangle^2$$

$$\Delta^2 = \langle a-a', v\rangle^2 + \langle a-a', v^\perp\rangle^2$$

$$\hat{\mu}(a) = \langle a, v\rangle$$
$$\hat{\mu}(a') = \langle a', v\rangle$$

$$E\left[\langle a-a', v\rangle^2\right] = \frac{\Delta^2}{d}$$

$$E\left[\langle a-a', v\rangle \cdot \sqrt{d}\right] = \Delta$$

for $i = 1$ to k

$\hat{v}_i \sim G_d(0, 1)$

$v_i = \frac{\hat{v}_i}{\|\hat{v}_i\|} \cdot \frac{\sqrt{d}}{\sqrt{k}}$

Not Data Adaptive!

for points $a_j \in A$

for $i \in [k]$

$\mu(a_j)_i = \langle a_j, v_i \rangle$

$a_j \to \mu(a_j) = (\mu(a_j)_1, \mu(a_j)_2, \ldots, \mu(a_j)_k)$

$(1 \pm \varepsilon)$ error w.p. $1 - \delta$

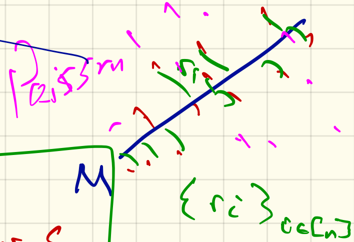
$\hookrightarrow k = O\left(\frac{1}{\varepsilon^2} \log \frac{n}{\delta}\right)$

1% $\Rightarrow$ 10,000

works for about

$k = 500 - 10{,}000$

Noise in Data



- Spurrious Readings : $\boxed{\text{outliers}}$

  ↳ adversarial examples, data poisoning

- Measurement Error : Gaussian Noise

  benign, SSE formulations, regularization

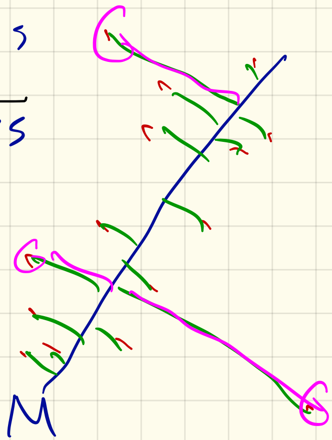
- Background Data : mixed in, or $\boxed{\text{missing}}$

  Matrix completion

Outliers

Input $X \in \mathbb{R}^{n \times d}$

- Build Model, Remove Outliers
- Density-based Approaches



1. Build model M on X
2. For each $x \in X$, find residual

   $r_x = d(M(x), x)$

3. For $x \in X$, with r_x too large $\rightarrow$ remove from X.

   $\hookrightarrow \tilde{X}$

4. Go back to Step 1. w/ $\tilde{X}$

1. threshold
2. percentage
   (remove 10%)

percentage of X
not $\tilde{X}$

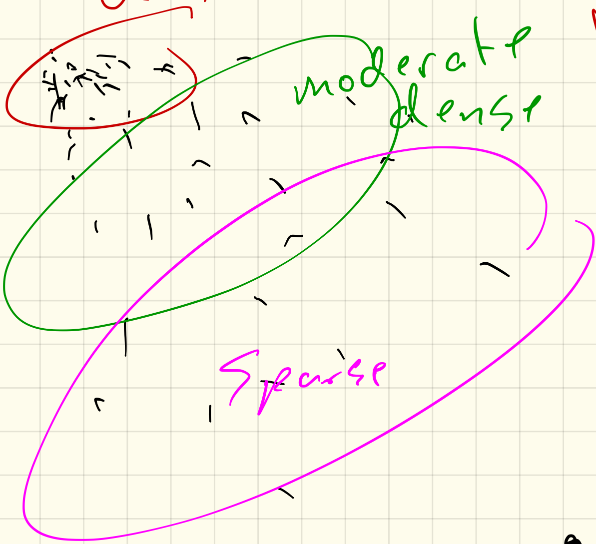
Density-based Approach

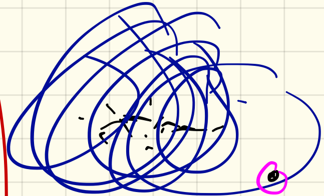
- DB-Scan Clustering.

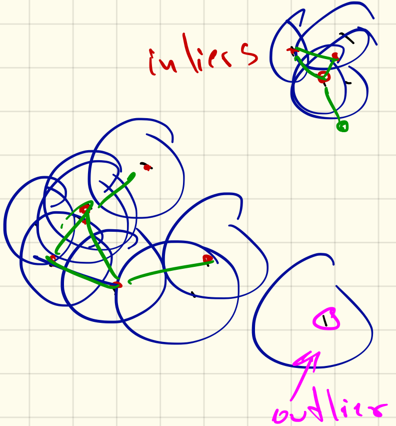
two param r , T

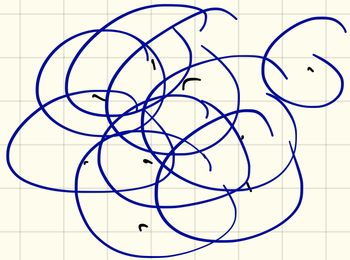
radius threshold

Very dense

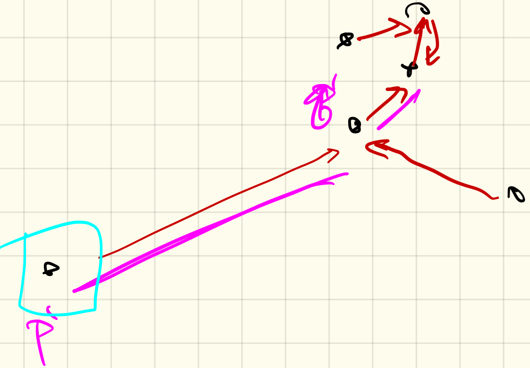


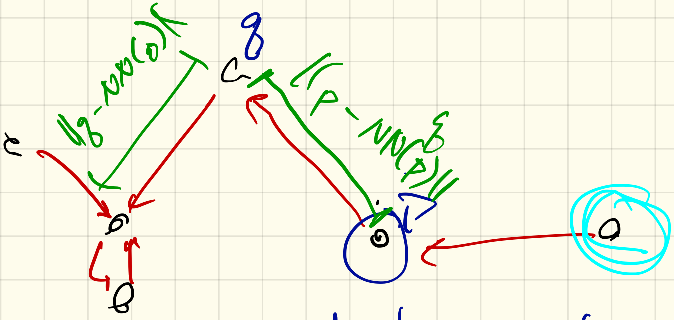






Reverse-Nearest Neighbor





q

distance to NN(p)

$\|q-p\|$

compare to

$\|q-NN(q)\|$

If comparable

p <u>not</u> outlier

Matrix Completion

Input $A \in \mathbb{R}^{n \times d}$ but w/ some ? x

$$A = \begin{bmatrix} 3 & 4 & \boxed{x} & 8 \\ 1 & \boxed{x} & 5 & \boxed{x} \\ \boxed{x} & 4 & \boxed{x} & 9 \\ 9 & 7 & 1 & \boxed{x} \\ 2 & 2 & \boxed{x} & \boxed{x} \end{bmatrix}$$

mask $\Omega = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$

5.125

replace each x by average of row and column.

$\frac{3+1+9+2}{4} = 3.75$ $\quad \frac{4+9}{2} = 6.5 \quad \Longrightarrow \quad \frac{10.25}{2} = 5.125$

assume A low-rank (w/ noise)

$$\phi_\lambda(S) = \mathrm{diag}\left((s_{11}-\lambda)_+, (s_{22}-\lambda)_+, \ldots, (s_{dd}-\lambda)_+\right)$$

$$(x)_+ = \max\{x, 0\}$$

$$A^* = \underset{X \in \mathbb{R}^{n \times d}}{\arg\min} \frac{1}{2} \| \Pi_\Omega (A - X) \|_F^2 + \lambda \|X\|_*$$

nuclear norm

$\|\|$

sum singular values

Matrix Completion

Initialize $X_{ij} \leftarrow \begin{cases} \Pi_\Omega(A_{ij}) & \text{if } (i,j) \in \Omega \\ \text{ave}(\text{rows } i, \text{col } j) & \text{o.w.} \end{cases}$

repeat

$U S V^T \leftarrow \mathrm{svd}(X)$

$\hat{X} \leftarrow U \phi_\lambda(S) V^T$

$X \leftarrow \Pi_\Omega(A) + \Pi_\Omega^\perp(\hat{X})$ ← replace known elements.

return $\hat{X}$