28.02.2020

$$A = \sum_{i=1}^{d} \delta_i u_i v_i^T \qquad rank(B) \leq k \leq d$$

$\|A\|_F \qquad \|A\|_2 = \max_{\|v\| \leq 1} \|Av\|$ (Spectral norm)

$$* \ \|A\|_F^2 = \sum_{i=1}^{d} \delta_i^2$$

$$a_j = \sum_{i=1}^{d} (a_j v_i) v_i^T \Rightarrow A = \sum_{i=1}^{d} (A v_i) v^T$$

row

$$\|A\|_F^2 = \sum_j |a_j|^2 = \sum_j \sum_{i=1}^{d} (a_j v_i)^2 = \sum_{i=1}^{d} |A v_i|^2 = \sum_{i=1}^{d} \delta_i^2$$

$$A_k = \sum_{i=1}^{k} \delta_i u_i v_i^T \qquad \|A - A_k\|_F^2 = \sum_{i=k+1}^{d} \delta_i^2$$

$* \ \|A - A_k\| \leq \|A - B\|_F$ * The rows of A_k are the projections of the rows of A onto the V_k.

$$a_j \perp V_k = \sum_{i=1}^{k} (A v_i) v_i^T = \sum_{i=1}^{k} \delta_i u_i v_i^T = A_k$$

Let $B = \underset{rank(C) \leq k}{\operatorname{argmin}} (\|A - C\|_F)$, $V = span(B)$ $\quad dim(V) \leq K$

Reduces $\|A - C\|_F$

$$\|A - A_k\| \leq \|A - B\|$$

$\downarrow$ best fit V_k $\qquad \searrow V$

$T.V \longrightarrow V$

Rank Nullity Thm:

$$\underbrace{dim(Im(T))}_{= rank(T)} + dim(Ker(T)) = dim(V)$$

$$* \quad \|A - A_k\|_2^2 = \delta_{k+1}^2 \qquad \|A - A_k\|_2 = \|A - A_k\|_F$$

$$A = \sum_{i=1}^{d} \delta_i u_i v_i^T \qquad A_k = \sum_{i=1}^{k} \delta_i u_i v_i^T$$

$$(A - A_k) = \sum_{i=k+1}^{d} \delta_i u_i v_i^T \qquad v = \sum_{i=1}^{d} c_i v_i \qquad \sum_{i=1}^{d} c_i^2 = 1$$

$$\|A - A_k\|^2 = \max_{\|v\|=1} |(A - A_k)v| = \max_{\|v\|=1} \left| \sum_{i=k+1}^{d} \delta_i u_i v_i^T \sum_{j=1}^{d} c_j v_j \right|.$$

$$= \max_{\|v\|=1} \left| \sum_{i=k+1}^{d} c_i \delta_i u_i v_i^T v_i \right| = \max_{\|v\|=1} \left| \sum_{i=k+1}^{d} c_i \delta_i u_i \right|$$

$$= \sqrt{\sum_{i=k+1}^{d} c_i^2 \delta_i^2}$$

$$\|A - A_k\|_2 \leq \|A - B\|_2$$

$$\|A - A_k\|_2^2 = \delta_{k+1}^2 \qquad Null(B) = \{v : Bv = 0\} \Rightarrow \dim(Null(B)) \geq d - k \quad \left(\begin{matrix}\text{by} \\ \text{Rank} \\ \text{Nullity} \\ \text{Thm}\end{matrix}\right)$$

$$0 \neq z \in Ker(B) \wedge span\{v_1, \ldots, v_{k+1}\}$$

$$\underbrace{\|A - B\|_2^2}_{\text{max by defn.}} \geq |(A - B)z|^2 \qquad Bz = 0, \ \|A - B\|_2^2 \geq |Az|^2$$

$$|Az|^2 = \left| \sum_{i=1}^{d} c_i u_i v_i^T z \right|^2 = \sum_{i=1}^{k+1} \delta_i^2 (v_i^T z)^2 \geq \delta_{k+1}^2 \underbrace{\sum_{i=1}^{k+1} (v_i^T z)^2}_{=1}$$

$$\geq \delta_{k+1}^2 = \|A - A_k\|_2^2$$

$*\ \|A_k - B\|_F^2 \le 5k \cdot \|A-B\|_2^2 \qquad \text{rank}(B) = k$

Let $\{\underbrace{v_1, \dots, v_k, u_1, \dots, u_k}_{u_1, \dots, u_p}\} \xrightarrow{\text{span}} \text{span}(A_k, B)$

$p \le 2k$

$$\left.\begin{array}{ll} u_i = v_i & 1 \le i \le k \\ u_i = u_i & k+1 \le i \le p. \end{array}\right\} \text{ extended variables}$$

$$\|A_k - B\|_F^2 = \sum_{i=1}^{k} |(A_k - B)u_i|^2 + \sum_{i=k+1}^{p} |(A_k - B)u_i|^2$$

$$= \sum_{i=1}^{k} |(A_k - B)u_i|^2 + \sum_{i=k+1}^{p} |Bu_i|^2$$

$$\le k\|A-B\|_2^2 + \sum_{i=k+1}^{p} |Au_i + (B-A)u_i|^2$$

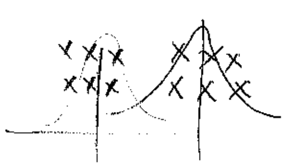
$$\le k\|A-B\|_2^2 + 2\underline{\sum_{i=k+1}^{p} |Au_i|^2} + 2\sum_{i=k+1}^{p} |(B-A)u_i|^2$$

$$\le k\|A-B\|_2^2 + 2k\,\sigma_{k+1}^2(A) + 2k\|A-B\|_2^2$$

$$\le k\|A-B\|_2^2 + 2k\|A-B\|_2^2 + 2k\|A-B\|_2^2$$

$$= 5k\|A-B\|_2^2$$

- Clustering is useful when the points are separable (in terms of mean, variance).
- When the distributions are overlapped the machine learner will classify wrong.

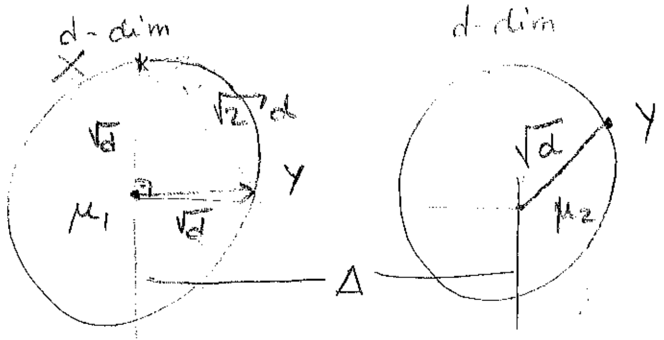


We can cluster <u>when</u> the means are far apart and variances do <u>not</u> overlap much.

$$\underset{(\mu_1, \sigma_1)}{\overset{(d\text{-dim})}{F}} = w_1 F_1 + w_2 \overset{(d\text{-dim})}{F_2} \quad \sim \quad w_1 + w_2 = 1 \qquad 0 \leq w_i \leq 1$$

$$(\mu_1, \sigma_1) \qquad (\mu_2, \sigma_2)$$

What's the structure of data to make us use ML?



Where will Y most likely be? Since the mass is located in equator, it's likely to be on the circle which is orthogonal to X's direction.
(north pole)

$$X = (\sqrt{d} + O(1), O(1), \ldots O(1))$$

$$Y = (O(1), \ldots, \sqrt{d} + O(1), O(1))$$

$$|X - Y|^2 \sim 2d \pm O(\sqrt{d})$$

$$|X - Y|^2 \sim \Delta^2 + 2d \pm O(\sqrt{d})$$

$$\Rightarrow 2d \pm O(\sqrt{d}) \leq \Delta^2 + 2d \pm O(\sqrt{d}) \qquad (*)$$

The max. of LHS $\leq$ the min of RHS of $(*)$

$$\cancel{2d} + O(\sqrt{d}) \leq \Delta^2 + \cancel{2d} - O(\sqrt{d})$$

$$2 \cdot O(\sqrt{d}) \leq \Delta^2$$

$$\Delta \sim \Omega(d^{1/4})$$

Why not random proj.? Not only the data points will be randomly projected but also the distances will be randomly projected