16 Feb 2022

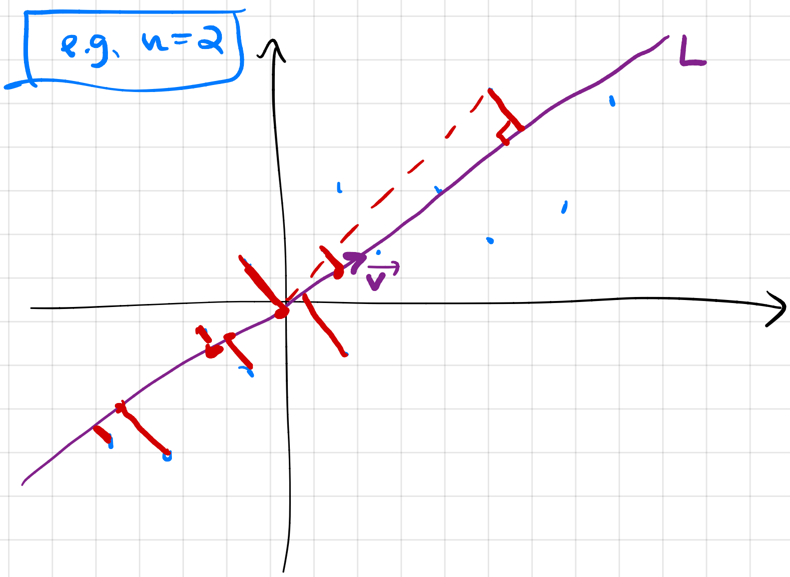
Singular Value Decomposition

Think of data points $a_1, a_2, \ldots, a_m \in \mathbb{R}^n$.

Organize them as rows of a matrix

$$A = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \qquad A \in \mathbb{R}^{m \times n}.$$

Plot rows of A as a "point cloud" in $\mathbb{R}^n$:



Q: Is there a line through $\vec{0}$ that is fairly close to all points?

Look for a unit vector $\vec{v}$ s.t. $\|\vec{v}\|_2 = 1$

This defines a line

$$L = \{ \alpha \cdot \vec{v} \mid \alpha \in \mathbb{R} \}.$$

Try to find a point $\alpha_i \cdot v$ on L close to each data pt a_i.

Measure goodness of fit by $\sum_{i=1}^{m} \|a_i - \alpha_i v\|^2$.

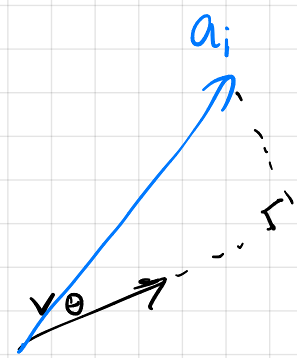
Pythagoras: $\forall i \quad \|a_i\|_2^2 = \|\alpha_i v\|_2^2 + \|a_i - \alpha_i v\|_2^2$

$$\sum_{i=1}^{m} \|a_i\|_2^2 = \sum_{i=1}^{m} \alpha_i^2 + \sum_{i=1}^{m} \|a_i - \alpha_i v\|_2^2$$

constant depending on data but not on $\vec{v}$.

Loss that we want to minimize

Solve loss minimization by maximizing $\sum_{i=1}^{m} \alpha_i^2$.



α_i = length of projection of a_i on the line determined by unit vector $\vec{v}$.

$= \|a_i\|_2 \cdot \cos\theta$

$= \|a_i\|_2 \|v\|_2 \cos\theta$

$= \langle a_i, v \rangle$.

Least squares problem is reduced to finding $\vec{v}$ s.t. $\|v\|_2 = 1$, maximizing $\sum_{i=1}^{m} \langle a_i, v \rangle^2$,

$\mathbin{/\!/}$

$\|Av\|_2^2$.

Approximating point cloud with a k-dimensional linear subspace?

Could try following same plan: choosing k linearly independent unit vectors $v_1, \ldots, v_k$

Approximate each point a_i with $\alpha_{i1} v_1 + \alpha_{i2} v_2 + \cdots \alpha_{ik} v_k$.

Minimize $\sum_{i=1}^{m} \| a_i - (\alpha_{i1} v_1 + \ldots + \alpha_{ik} v_k) \|_2^2$.

This is again equivalent to maximizing

$$\sum_{i=1}^{m} \| \alpha_{i1} v_1 + \cdots + \alpha_{ik} v_k \|_2^2$$

If $v_1, \ldots, v_k$ are <u>orthonormal</u> meaning

$\|v_i\|_2^2 = 1 \quad \forall i$ and $\langle v_i, v_j \rangle = 0$ for $i \neq j$

then $\alpha_{ij} = \langle a_i, v_j \rangle$ and we're

minimizing $$\sum_{i=1}^{m} \sum_{j=1}^{k} \langle a_i, v_j \rangle^2.$$

Given A, find an orthonormal set of vectors

$v_1, \ldots, v_k \in \mathbb{R}^n$ to maximize $\sum_{i=1}^{k} \sum_{j=1}^{k} \langle a_i, v_j \rangle^2.$

$$V = \begin{pmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_k \\ | & | & & | \end{pmatrix} \qquad = \| AV \|_F^2.$$

$$\|M\|_F \triangleq \left(\sum_{i,j} M_{ij}^2 \right)^{1/2}.$$

Theme of today + Friday...

The greedy algorithm solves this problem!

Meaning: the optimal $v_1, \ldots, v_k$ can be found by solving

1. $v_1 \in \arg\max \{ \|Av\|_2^2 \mid \|v\|_2 = 1 \}$
2. $v_2 \in \arg\max \{ \|Av\|_2^2 \mid \|v\|_2 = 1, \langle v_1, v \rangle = 0 \}$

$\vdots$

k. $v_k \in \arg\max \left\{ \|Av\|_2^2 \;\middle|\; \|v\|_2 = 1,\ \langle v_i, v \rangle = 0 \ \forall i < k \right\}$