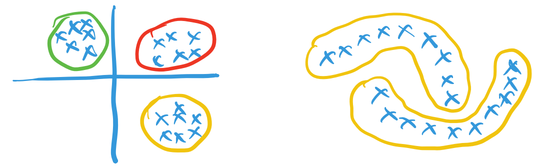
Machine Learning

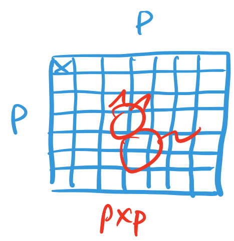
Supervised (x_i, y_i)

Unsupervised (x_i) → Clustering



1. Principal Component Analysis (PCA) ⟵ Dimensionality Reduction
2. k-Means Clustering

Image Data:

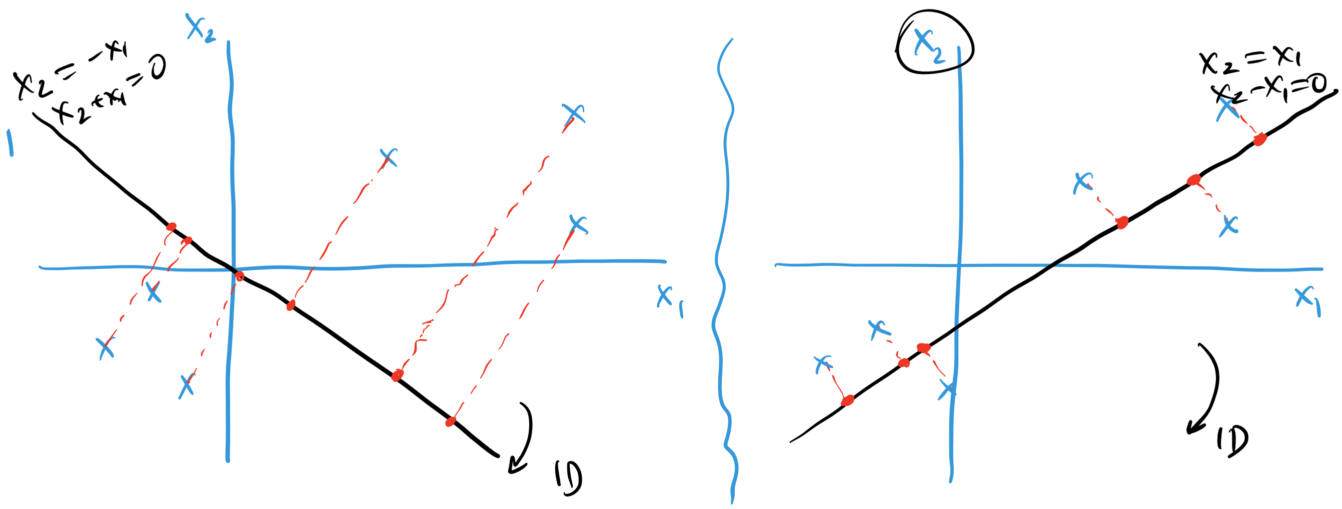


each grid is a pixel
→ between [0,1]

$x_i \in \mathbb{R}^{p^2 \times 1}$

MNIST : # Features > 500

PCA Motivation (2D problem)



Variance is lower

spread of points is linear

p-dim dataset

↓

Compute p principal components which are the axes on

variance

$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

which variance is maximised

↓

pick subset, $m < p$ to summarise data

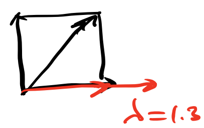
Eigenvalues and Eigenvectors of Matrices

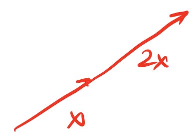
$A \in \mathbb{R}^{m \times n}$ $\quad x \in \mathbb{R}^n \quad Ax = b,\ b \in \mathbb{R}^m$

view a matrix A as a function/map that maps vectors in $\mathbb{R}^n$ to vectors in $\mathbb{R}^m$

Eigenvector x of matrix A satisfies:

$$Ax = \lambda x \quad , \quad (\lambda, x) \text{ eigenpair}$$





$$A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix} \qquad Ax = \lambda x \Rightarrow Ax - \lambda x = 0$$

$$\Rightarrow (A - \lambda I)\lambda = 0$$

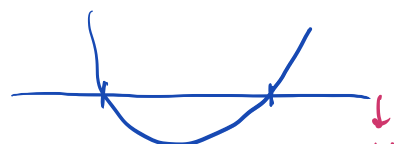
$$\Rightarrow \det(A - \lambda I) = 0$$

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \underline{ad - bc}$$

$$\det(A - \lambda I) = \det\left(\begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}\right)$$

$$= \det\begin{pmatrix} -\lambda & 1 \\ -2 & -3-\lambda \end{pmatrix}$$

$$= \lambda(3+\lambda) + 2 = \lambda^2 + 3\lambda + 2$$



$$Ax_1 = \begin{pmatrix} -2 \\ 4 \end{pmatrix} = -2\begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Solve for λ satisfying $\boxed{\lambda^2 + 3\lambda + 2 = 0}$ — quadratic

$$\lambda_1 = -2 \quad , \quad \lambda_2 = -1$$

usually, order eigenvalues in decreasing size of absolute value

Eigenvector of $\lambda_1 = -2$

$$Ax_1 = -2x_1$$

$$\Rightarrow \begin{pmatrix} 0 & 1 \\ -2 & -3 \end{pmatrix}\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix} = -2\begin{pmatrix} x_{11} \\ x_{12} \end{pmatrix}$$

$$\boxed{x_{12} = -2x_{11}} - (1)$$

$$-2x_1 - 3x_{12} = -2x_{12} - (2)$$

$$x_1 = \begin{pmatrix} 1 \\ -2 \end{pmatrix} \quad x_1 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

$$x_1 = \begin{pmatrix} 2 \\ -4 \end{pmatrix} \quad x_1 = \begin{pmatrix} 1/\sqrt{5} \\ -2/\sqrt{5} \end{pmatrix}$$

eigenvector is the norm-1 vector satisfying the requirement: $\sqrt{1^2 + (-2)^2}$

$= \sqrt{5}$

$x_i \to \alpha_1^T x_i$:
$\frac{1}{N}\sum_{i=1}^{N} \alpha_1^T x_i = \alpha_1^T \frac{1}{N} \sum x_i$
$= \alpha_1^T \bar{x}_n$

PCA

- sample of N data points: $x_{1}, \dots, x_{N}$ $\quad x_i \in \mathbb{R}^p$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Want to find the first PC. Project the data from $\mathbb{R}^p \to \mathbb{R}^1$
  - $\alpha_1 = \begin{bmatrix} \alpha_{11} \\ \alpha_{12} \\ \vdots \\ \alpha_{1p} \end{bmatrix} \in \mathbb{R}^p$ $\qquad \alpha_1^T x_i \in \mathbb{R}^1$
  - Want variance of projected data to be maximised

$$\frac{1}{N-1} \sum_{i=1}^{n} (\underline{\alpha_1^T x_i} - \alpha_1^T \bar{x})^2 = \alpha_1^T \left(\frac{1}{N-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T \right) \alpha_1$$

$$= \alpha_1^T S \alpha_1$$

Sample Covariance Matrix

Find $\alpha_1 \in \mathbb{R}^3$ s.t. $\alpha_1^T S \alpha_1$ maximised where S = sample covariance matrix of $[x_i]$

- constraint: $\alpha_1^T \alpha_1 = 1 \to \|\alpha_1\| = 1$

$$\max_{\alpha_1} \alpha_1^T S \alpha_1 \quad \text{s.t.} \quad \alpha_1^T \alpha_1 = 1 \to \alpha_1^T \alpha_1 - 1 = 0$$

Lagrangian: $L(\alpha, \lambda) = \underline{\alpha_1^T S \alpha_1} - \lambda(\alpha_1^T \alpha_1 - 1)$

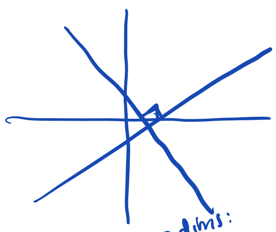
$$\frac{\partial L}{\partial \alpha_1} = 2 S \alpha_1 - 2 \lambda \alpha_1 = 0$$

$$\boxed{S \alpha_1 = \lambda \alpha_1} \longrightarrow$$ α_1 is an eigenvector of S with eigenvalue λ

$$\max \alpha_1^T S \alpha_1 = \alpha_1^T \lambda \alpha_1 = \lambda \underbrace{\alpha_1^T \alpha_1} = \lambda = \lambda_1$$

<u>Second PC?</u> $\alpha_2 \in \mathbb{R}^p$ where α_2 maximises $\underline{\alpha_2^T S \alpha_2}$ s.t. $\underline{\alpha_2^T \alpha_2} = 1$

Want second direction to be orthogonal to first direction $\underline{\alpha_1^T}\alpha_2 = 0$



$$L(\alpha_2, \lambda, \phi) = \alpha_2^T S \alpha_2 - \lambda(\alpha_2^T \alpha_2 - 1) - \phi(\alpha_1^T \alpha_2)$$

$$\frac{\partial L}{\partial \alpha_2} = \alpha_1^T [2 S \alpha_2 - 2\lambda \alpha_2 - \phi \alpha_1] = 0$$

$$2 S\alpha_2 - 2\lambda\alpha_2 = 0$$

$$\boxed{S\alpha_2 = \lambda \alpha_2}$$

α_2 is an eigenvector of S with eigenvalue λ_2

$$\Rightarrow 2\alpha_1^T S \alpha_2 - 2\lambda \alpha_1^T \alpha_2 - \phi \alpha_1^T \alpha_1 = 0$$

$$\Rightarrow 2\alpha_2^T [S\alpha_1] - 2\lambda\alpha_1^T\alpha_2 - \phi\alpha_1^T\alpha_1 = 0$$

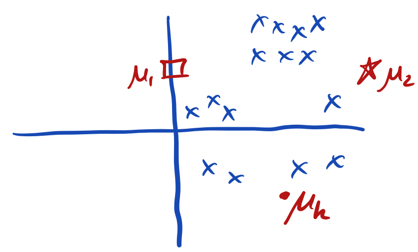
$$\Rightarrow \underbrace{2\alpha_2^T \lambda_1 \alpha_2}_{=0} - 2\lambda \underbrace{\alpha_1^T \alpha_2}_{=0} \boxed{-\phi \alpha_1^T \alpha_1} = 0$$

$$-\phi \underbrace{\alpha_1^T \alpha_1}_{=1} = 0$$

$$\phi = 0$$

$$\alpha_2^T S \alpha_2 = \alpha_2^T \lambda \alpha_2 = \lambda_2$$

k-means: (k)



1. Pick initial centroids: $\mu_1, \dots, \mu_k$
2. Repeat until cluster centers don't change

(Assignment) For each i, $c_i = \arg\min_j \|x_i - \mu_j\|_p$

(Recompute cluster centers) For each j, set

$$\mu_j = \frac{\sum_{i=1}^{n} 1\{c_i = j\} x_i}{\sum_{i=1}^{n} 1\{c_i = j\}}$$

= <u>average</u> of x_i's in cluster j