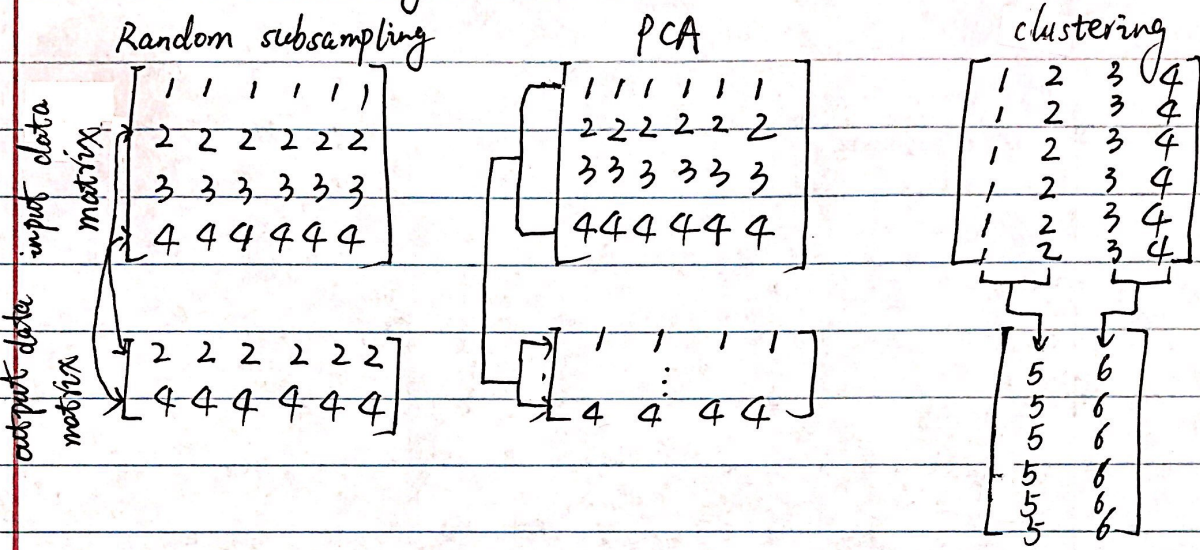
Dimention reduction techniques.

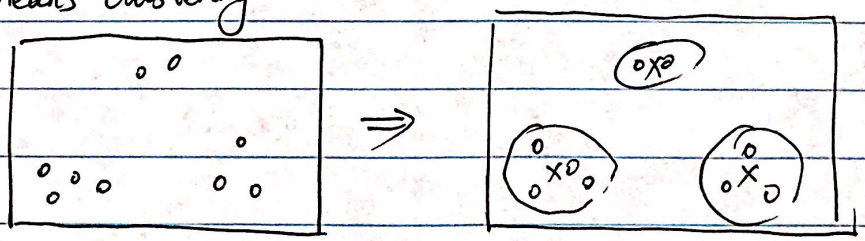
- Random subsampling
- PCA
- K-means clustering



1. Random subsampling

Given a set of P points we keep a random subsampling of $S < P$ of the entire set.

2. K-means clustering



Denoting by $\vec{c}_k$ the centroid of the kth cluster and S_k the set of indices of the subset of those P data points, denoted $\bar{x}_1 \dots \bar{x}_p$, belonging to this cluster, the desire that points in the kth cluster should lie close to its centriod may be written as.

$$\bar{c}_k \approx \bar{x}_p \quad \text{for all } p \in S_k.$$

for all $k = 1, \dots K$.

Centroid matrix: $\bar{C} = [\bar{c}_1 \ \bar{c}_2 \cdots \bar{c}_K]$ (a $N \times K$ vector)

Denote by $\bar{e}_k$ the kth standard basis vector (a $K \times 1$ vector with a 1 in the kth slot and zeros elsewhere). we may write $\bar{C}\bar{e}_k = \bar{c}_k$

$\therefore \quad \bar{C}\bar{e}_k \approx \bar{x}_p$ for all $p \in S_k$

Data matrix $\bar{X} = [\bar{x}_1 \ \bar{x}_2 \ \cdots \bar{x}_p]$ and

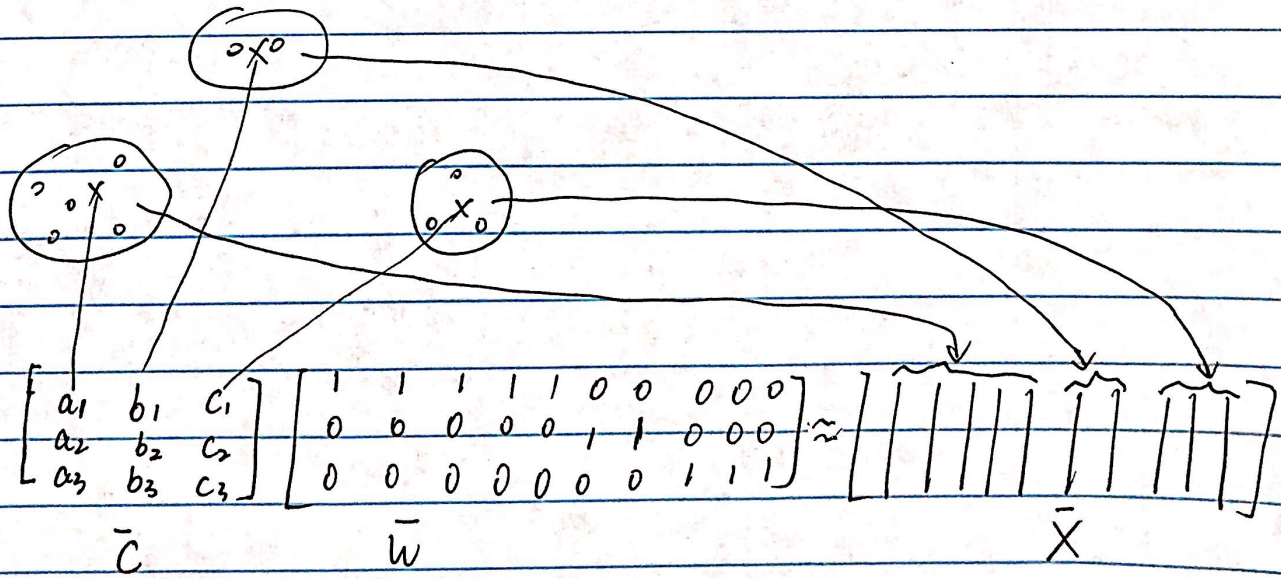
a $K \times P$ assignment matrix $\bar{W}$ ($\bar{w}_p = \bar{e}_k$ if $p \in S_k$, namely $\bar{C}\bar{w}_p \approx \bar{x}_p$ for all $p \in S_k$)

$\therefore \quad \bar{C}\bar{W} \approx \bar{X}$

$\therefore$ k-means optimization:

$$\underset{\bar{C}, \bar{W}}{\text{minimize}} \ \| \bar{C}\bar{W} - \bar{X} \|_F^2 \qquad \text{(Non-convex)}$$

subject to $\bar{w}_p \in \{\bar{e}_k\}_{k=1}^{K} \quad p = 1, \ldots P$



Algorithm: The k-means algorithm

Input(s): Data matrix $\bar{X}$, centroid matrix $\bar{C}$ initialized (e.g., randomly) and assignment matrix $\bar{X}$ initialized at zero

Output(s): Optimal centroid matrix $\bar{C}^*$ and assignment matrix $\bar{W}^*$

Repeat until convergence: (e.g., until C does not change)

(1) update $\bar{w}$ (assign each data point to its closest centroid)

for $p = 1, \ldots, P$

set $\bar{w}_p = \bar{e}_{k^*}$ where $k^* = \underset{k=1,\ldots k}{\operatorname{argmin}} \| \bar{c}_k - \bar{x}_p \|_2^2$

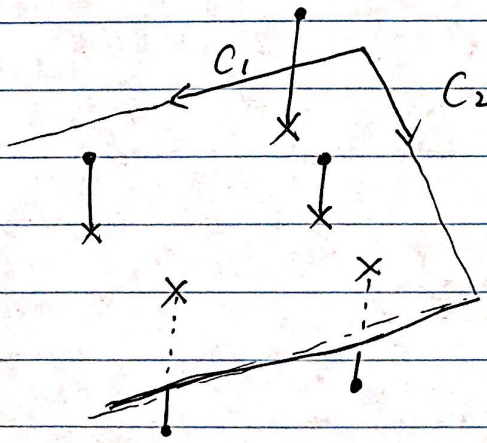
(2) Update $\bar{C}$ (assign each centroid the average of its current points)

for $k = 1, \ldots K$

Denote S_k the index set of points $\bar{x}_p$ currently assigned to the kth cluster Set $\bar{c}_k = \frac{1}{|S_k|} \sum_{p \in S_k} \bar{x}_p$

To overcome the issue of non-convexity of k-means in practice we usually run the algorithm multiple times with different initialization, seeking out the lowest possible minimum of the objective.

3. PCA (Principle component analysis)



PCA works by simply projecting the data onto a suitable lower dimentional feature subspace, that is one which hopefully preserves the essential geometry of the original data.

Suppose that we have P data points $\bar{x}_1 \ldots \bar{x}_p$, each of dimension N The goal with PCA is, for some user chosen dimentional $K < N$, to find a set of K vectors $\bar{c}_1, \bar{c}_2 \cdots \bar{c}_k$ that represent the data fairly well.

Put formally, we want for each $p = 1 \dots P$

$$\sum_{k=1}^{K} \bar{c}_k w_{kp} \approx \bar{x}_p$$

Stacking the desired spanning vectors column-wise into the $N \times K$ matrix $\bar{C}$ as $\bar{C} = [\bar{c}_1 | \bar{c}_2 | \dots | c_K]$ and denoting $\bar{w}_p = [w_{1,p}\ w_{2,p} \dots w_{K,p}]$ this can be written for each p as

$$\bar{C}\bar{w}_p \approx \bar{x}_p$$

Note: once $\bar{C}$ and $\bar{w}_p$ are learned the new k-dimentional feature representation of $\bar{x}_p$ is then the vector $\bar{w}_p$. By denoting $\bar{W} = [\bar{w}_1 | \bar{w}_2 | \dots \bar{w}_p]$ the $K \times P$ matrix of weights to learn, and $\bar{X} = [\bar{x}_1 | \bar{x}_2 | \dots \bar{x}_p]$ the $N \times P$ data matrix, all P of these can be written as

$$\bar{C}\bar{W} \approx \bar{X}$$