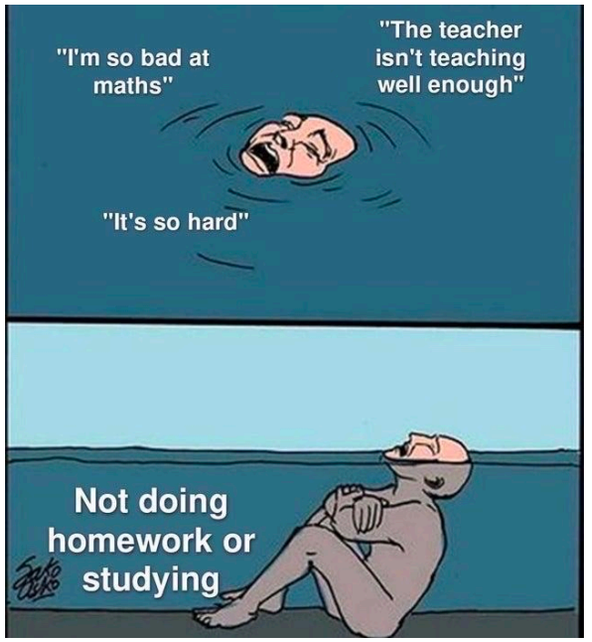
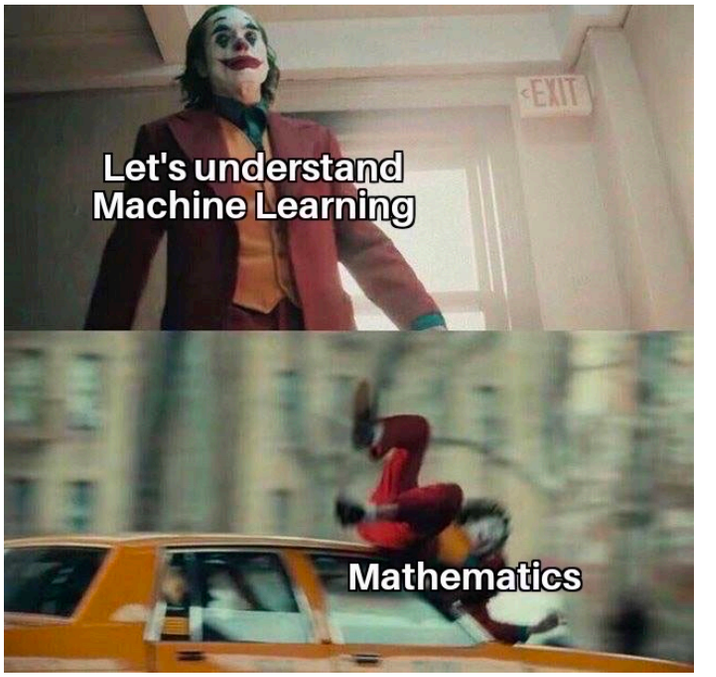
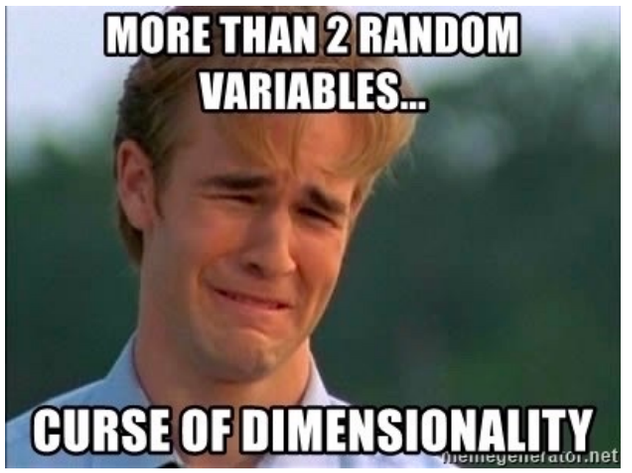
June 13, 2023.

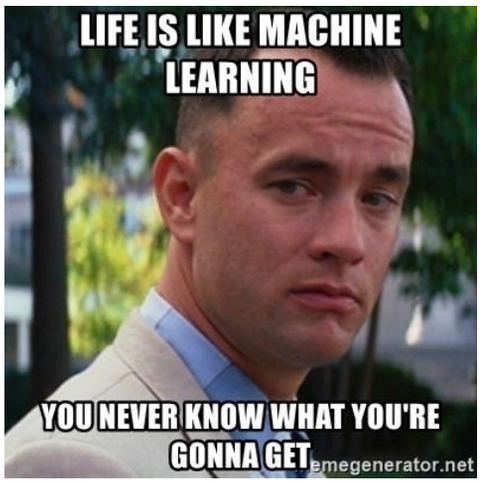
DSML : Math for ML.

Optimization 5: Principal Component Analysis.









Class begins @ 9:03 p.m. sharp.

Dimensionality Reduction.

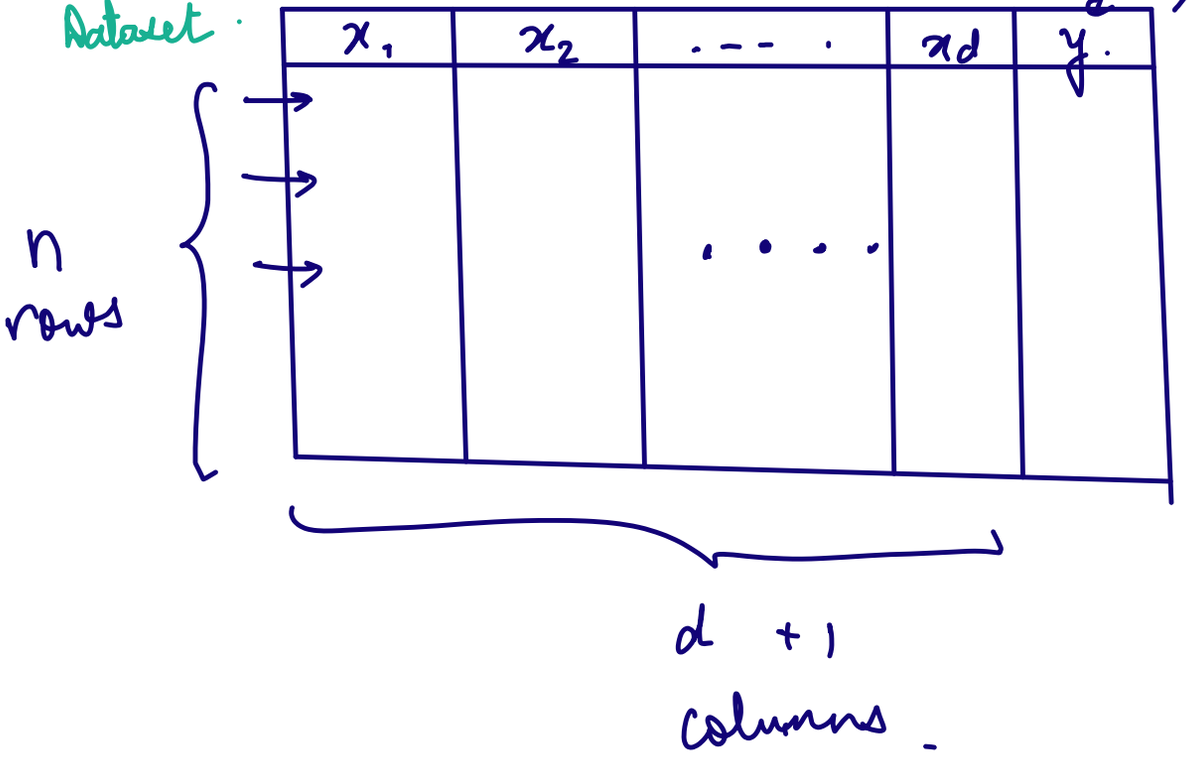
$d >> 4$

$$D = \{ (\bar{x}_i, y_i) : \bar{x}_i \in \mathbb{R}^d \}_{i=1}^{n}$$

Labelled Dataset.

target / label.

$$\bar{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$



Examples:

1] Principal component Analysis (PCA)

2] t-SNE.

3] UMAP.

Objective: ① Visualize this dataset (Reduce dimension to 2 or 3).

② Keep as much information as possible.

Why is Dimensionality reduction needed?

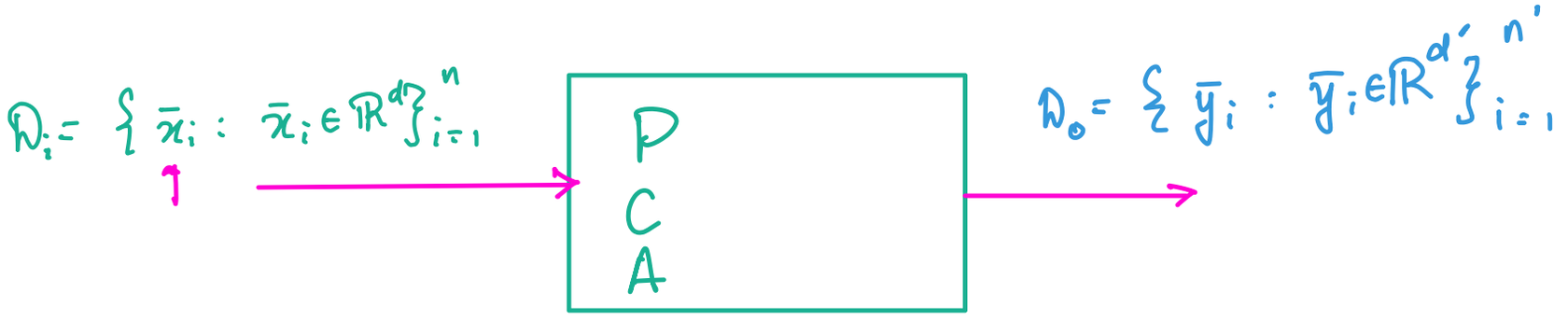
1] To make it easier to visualize high dimensional data.

2] less dimensions ⇒ lesser memory to store data

⇒ less computation while doing ML.

3] Makes it easier to code / implement ML algos.

PCA as a black box:



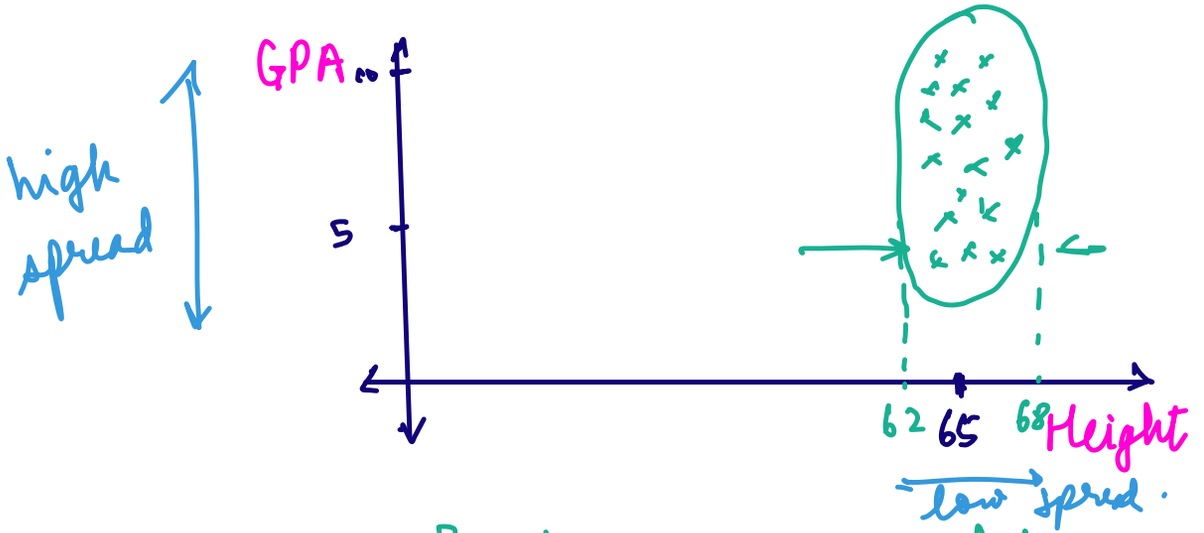
$$d \geq d'$$

$$n = n'$$

Examples.

1] GPA. (10) vs. Height.

Spread of values captures usefulness of the feature.

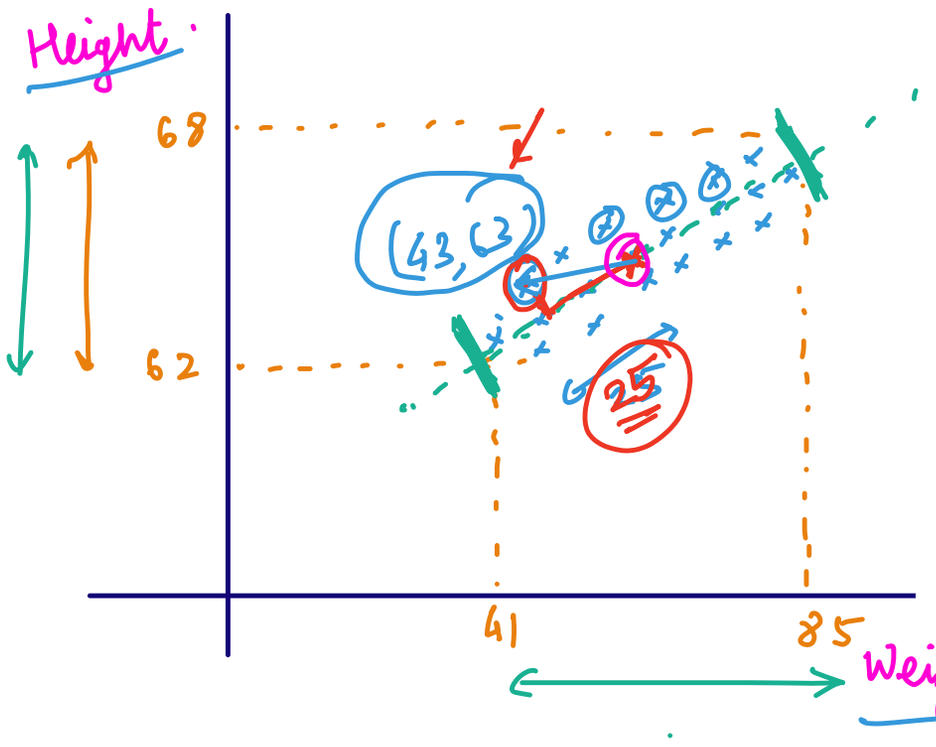


1] No correlation between Height & GPA.

2] Range of values is low.

3] The standard deviation / variance / spread of the height feature is low.

2] Height vs. Weight.



1] Calculate the centre of our data and change co-ordinate systems.

2] Standardize features

3] Find the direction of max. variance.

4] Project all points along this direction

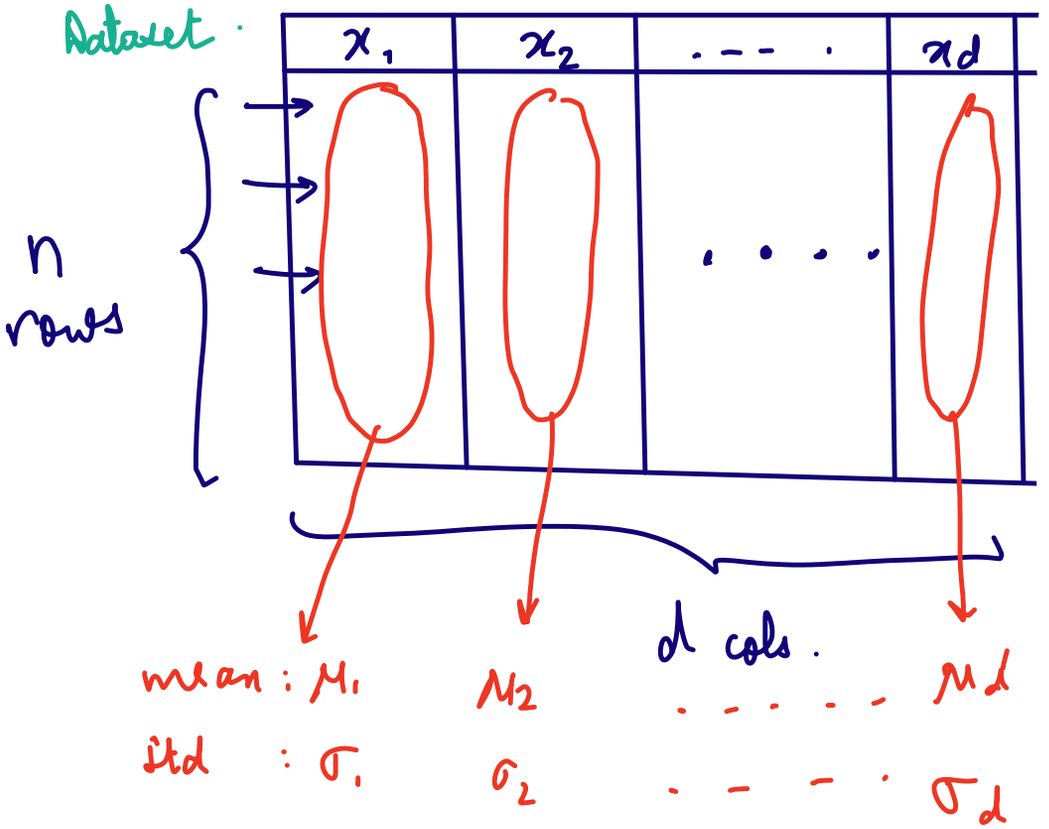
5] Keep doing steps 3 & 4 for d' iterations.

If we apply D.R. here, how to do it ?

Steps 1 & 2: Re-center & Standardize features. [Column standardization.]

$$\mathcal{D} = \{ \bar{x}_i : \bar{x}_i \in \mathbb{R}^d \}_{i=1}^{n}$$

Labelled Dataset.



1] Recenter:

means: $\mu_1, \mu_2, \ldots, \mu_d$.

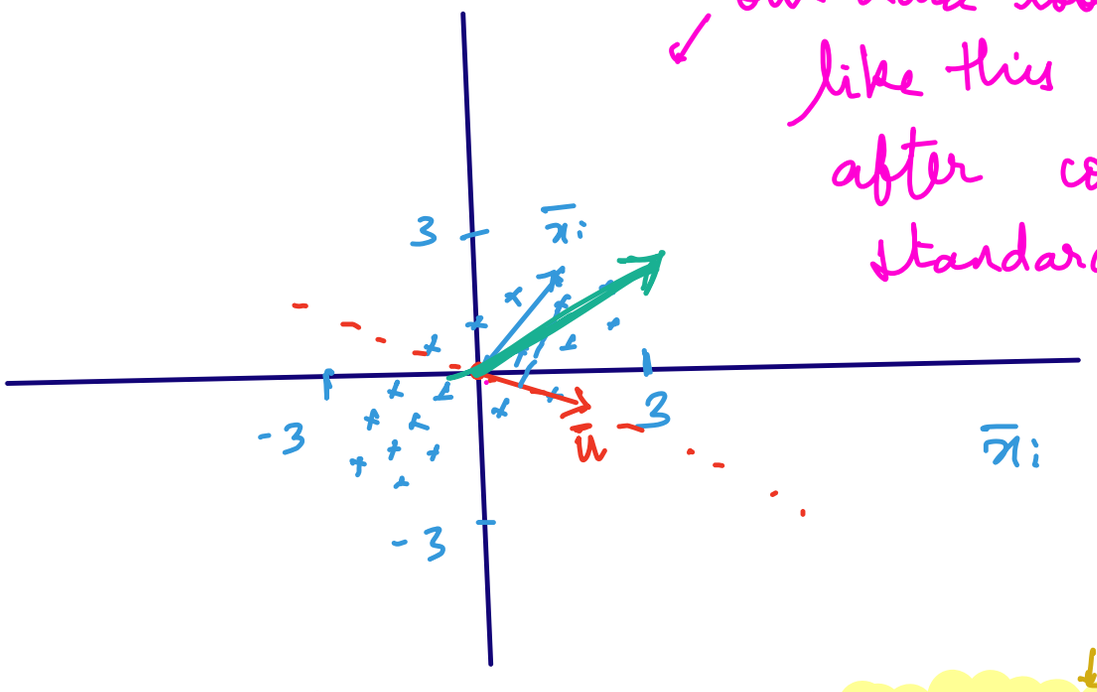
stds: $\sigma_1, \sigma_2, \ldots, \sigma_d$.

2] $$x_i' = \frac{x_i - \mu_i}{\sigma_i}$$

Apply this to all columns.

$$\mathcal{D}' = \{ \bar{x}_i' : \bar{x}_i' \in \mathbb{R}^d \}_{i=1}^{n}$$

Our data looks like this after column standardization.



Next, we will find the direction of max. variance.

→ Design our cost function.

Sum of Squares of Projection of $\bar{x}_i$ on $\bar{u}$: $\sum_{i=1}^{n} \left(\frac{\bar{x}_i^T \bar{u}}{\|\bar{u}\|} \right)^2$

will be more ↓ for better $\bar{u}$s

The optimization problem for PCA.

$$\max_{\bar{u}} \sum_{i=1}^{n} \left(\frac{\bar{x}_i^T \bar{u}}{\|\bar{u}\|} \right)^2$$

Let's assume $\bar{u}$ is a unit vector.

$$\max_{\bar{u}} \sum_{i=1}^{n} (\bar{x}_i^T \bar{u})^2$$

s.t. $\|\bar{u}\|^2 = 1$.

Lagrange multipliers.

$$\max_{\bar{u}} \sum_{i=1}^{n} (\bar{x}_i^T \bar{u})^2 + \lambda (\|\bar{u}\|^2 - 1)$$

$\|\bar{v}\|^2$ is same as $\sum_{i=1}^{n} (\bar{x}_i^T \bar{u})^2$ ← $f(\bar{x}_i ; \bar{u})$ $+ \lambda (\|\bar{u}\|^2 - 1)$

$\max_{\bar{u}}$

This is the problem we have to solve.

1] Take the gradient of $f(\bar{x}_i ; \bar{u})$ and set it to 0.

$$X\bar{u} = \bar{v}$$

$$D = \{ \bar{x}_i : \bar{x}_i \in \mathbb{R}^d \}_{i=1}^{n}$$

$X \rightarrow$ $\begin{bmatrix} \text{—} \ \bar{x}_1^T \ \text{—} \\ \text{—} \ \bar{x}_2^T \ \text{—} \\ \vdots \\ \text{—} \ \bar{x}_n^T \ \text{—} \end{bmatrix}_{n \times d}$ $\cdot$ $\bar{u}$ $\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_d \end{bmatrix}_{d \times 1}$ $=$ $\bar{v}$ $\downarrow$ $\begin{bmatrix} \bar{x}_1^T \bar{u} \\ \bar{x}_2^T \bar{u} \\ \vdots \\ \bar{x}_n^T \bar{u} \end{bmatrix}_{n \times 1}$

$\|\bar{v}\|^2$ is same as

$f(\bar{x}_i ; \bar{u})$

$$\max_{\bar{u}} \sum_{i=1}^{n} (\bar{x}_i^T \bar{u})^2 + \lambda (\|\bar{u}\|^2 - 1)$$

$$\Rightarrow \max_{\bar{u}} \bar{v}^T \bar{v} + \lambda (\|\bar{u}\|^2 - 1).$$

$$\Rightarrow \max_{\bar{u}} (X\bar{u})^T (X\bar{u}) + \lambda (\bar{u}^T \bar{u} - 1)$$

$$= \bar{u}^T X^T X \bar{u}$$

$$(A \cdot B)^T = B^T A^T$$

$$\max_{\bar{u}} \underbrace{\bar{u}^T x^T x \bar{u} + \lambda(\bar{u}^T\bar{u} - 1)}_{f(x;\bar{u})}.$$

Now, take gradient & set it to zero.

Let $f(\bar{x}) = \bar{x}^T A \bar{x}$

then $\nabla_{\bar{x}} f(\bar{x}) = (A + A^T)\bar{x}$

$\nabla_{\bar{u}}\ \bar{u}^T\bar{u} = 2\bar{u}$ $\quad A = x^T x$

$A^T = (x^T x)^T = x^T (x^T)^T = x^T x.$

$$\nabla_{\bar{u}} f(x;\bar{u}) = 2 \cdot x^T x \bar{u} + \lambda\, 2\bar{u}$$

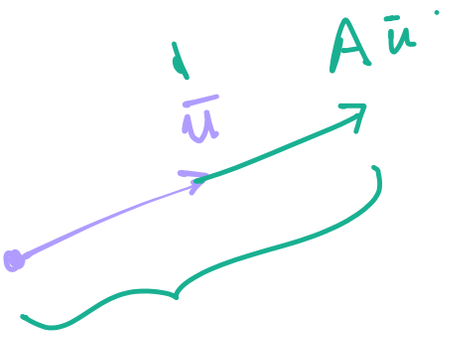
$$2\, x^T x \bar{u} + \lambda\, 2\bar{u} = 0$$

$$\boxed{x^T x\ \bar{u} = -\lambda\, \bar{u}}$$

$$X^T X \bar{u} = \underbrace{-\lambda}\, \bar{u}$$

$$\lambda' = -\lambda$$

$$\underbrace{X^T X}_{\text{matrix}} \bar{u} = \lambda' \bar{u}$$



$$\boxed{A \cdot \bar{u} = \lambda' \bar{u}}$$

constant.

$\bar{u}$ is an eigenvector of the matrix A with an eigenvalue of λ'.

Actually, $\bar{u}$ is the eigenvector with the maximum eigenvalue.

In general, our dataset has d-dim vectors.
If I want to get d' dimensions for dimensionality reduction, then:

1] Calculate the covariance matrix $X^T X$.

2] Next calculate the eigenvectors & eigenvalues of $X^T X$.

X $(n \times d)$

X^T $(d \times n)$

Here dimensions of $X^T X$ = $\underline{d \times d}$.

It is guaranteed that we will get d eigenvalues, eigenvector pairs.

Also, the d eigenvalues ≥ 0.

After step 2, we will have:

$$\{ (\bar{u}_1, \lambda_1), (\bar{u}_2, \lambda_2), (\bar{u}_3, \lambda_3) \cdots (\bar{u}_d, \lambda_d) \}$$

3] Now, sort the above so that:

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \cdots \geq \lambda_d .$$

4] Finally, my reduced dimension is obtained by selecting the top d' eigenvalue & eigenvector pairs.

$$\bar{y}_i = \begin{bmatrix} (\bar{x}_i^T \bar{u}_1) \\ (\bar{x}_i^T \bar{u}_2) \\ \vdots \\ (\bar{x}_i^T \bar{u}_{d'}) \end{bmatrix} \rightarrow \text{Final output.}$$

Practical Considerations while Using PCA.