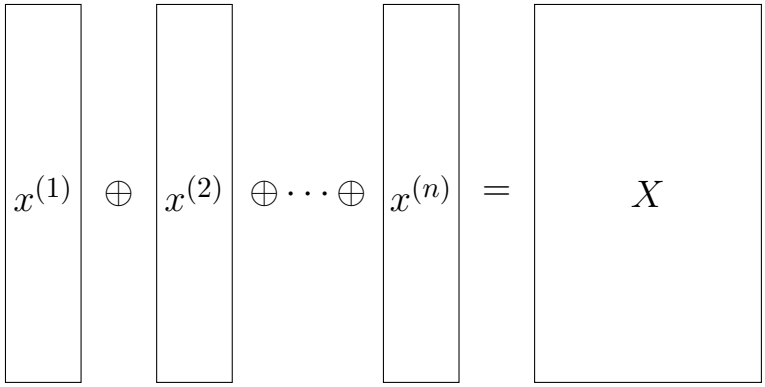
Chapter 2: Principal Component Analysis

John D Mangual

Machine learning is just linear regression and best fit lines. And that is just the Pythagorean theorem. Today let's discuss **Principal Component Analysis**.

Our data points are $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$ so we can merge them into a single matrix $X \in \mathbb{R}^{m \times n}$.



In Python is this done with the `np.hstack` or `np.vstack` commands (where `numpy` is shortened to `np`).

In beginning statistics we search for the best fit line. That means all our data points x approximately solve the same equation:

$$Ax + B \approx 0$$

What are the dimensions of A and B? Our vectors x have shape $1 \times n$ and the zero has shape 1×1, so that A has shape $n \times 1$ and B has shape 1×1. However, reduction to a line may lose much information.

What if we need more features? Let's take $n_0 < n$ feature.

What are the dimensions of A and B? Our vectors x have shape $1 \times n$ and the zero has shape $1 \times n_0$, so that A has shape $n \times n_0$ and B has shape $1 \times n_0$.

What is a reasonable numbe of data points? Some data sets brag as much as a billion, $m = 10^9$ data points. Each row of data could have as much as $n = 100$ features (or more) and we could like to reduce that to $n_0 = 10$.

If our data set included a few bit of inforation about each person:

- Height
- Favorite Ice Cream
- Birthday
- Occupation

Can we infer the **Birthday** and **Occupation** from the height and favorite ice cream flavor? Hopefully there is some type of correlation!

So we will take our data points X and find the covariance matrix. Hopefully our numbers do not have too much mistakes because they may propagage throughout our computation!

$$\text{Covariance} = X^T X \in \mathbb{R}^{m \times m}$$

What is the shape of our matrix? It is always okay to multiply a matrix with its own transpose:

$$(m \times n) \cdot (n \times m) = m \times m \quad \text{or} \quad (n \times m) \cdot (m \times n) = n \times n$$

n is the **number of features** and m is the **number of data points**, so we are doing the one on the right.

Next we decide which features are the most important. Using the eigenvalue decomposition:

$$X^T X \sim \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

Our matrix is not only square, but they are symmetric! So the eigenvalues $\lambda \in \mathbb{R}$ can be compared.

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n$$

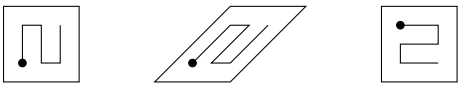
An example of a matrix whose eigenvalues cannot be a real number is a **rotation** by $90°$:

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \sim \begin{bmatrix} \sqrt{-1} & 0 \\ 0 & \sqrt{-1} \end{bmatrix}$$

Here is a 2×2 matrix which is not diagonal. There is only one eigenvector:

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

This is called a **shear**.



Our covariance matrix is symmetric: $(XX^T)^T = X^T(X^T)^T = XX^T$, so out our eigenvalues, our **features** are real numbers. In practice these features will be linear combinations and therefore totally ridiculous:

$$\frac{1}{2} \times \text{Height} + 2 \times \text{Birthday}$$

Let us proceed with our analysis of Principal Component Analysis.