Chapter 9: Exploratory Data Analysis (EDA)

- dot product: $\langle U, V\rangle = U_1V_1 + U_2V_2 + \dots +$
- norm: $\|V\| = \sqrt{\langle V, V\rangle}$
- unit vector: $\|V\| = 1$
- orthogonal: $\langle V, U\rangle = 0$

- projection $= \frac{\langle x, u_1\rangle}{\langle u_1, u_1\rangle} \cdot u_1 + \frac{\langle x, u_2\rangle}{\langle u_2, u_2\rangle} \cdot u_2 + \dots$

First check that, u_1 & u_2 are orthogonal

- PCA
- Transpose: $d_1 \times d_2$ and $d_2 \times d_1$
- Square matrix: $A' = A$
- $A.U = \lambda.U$, U = eigenvector and λ = eigenvalue

- Centroid $(\bar{x}) = \frac{1}{n}(x_1 + \dots + x_n)$ } Centering
- Replace each x_i by $x_i - \bar{x}$
- Normalization: $\dfrac{\text{values}}{\text{standard deviation}}$ ($s.d = \sqrt{\text{variance}}$)

Variance $= \frac{1}{n}(x_1^2 + x_2^2 + \dots)$

- subspace spanned by u
- when the variance along u is large indicate that the projection along u are interesting.
- Maximizing variance
- let u a unit vector
  - The projection of x along vector u is $\langle x, u\rangle . u$ (coefficient)
  - The variance is

$$\frac{1}{N}\sum_{i=1}^{N} \langle x_i, u\rangle^2$$

  - N times the variance is

$$\sum_{i=1}^{N} \langle x_i, u\rangle^2 = \sum_{i=1}^{N} u' x_i (x_i)' u = u' \left(\sum_{i=1}^{N} x_i (x_i)'\right) u$$

so, $S = \sum_{i=1}^{N} x_i (x_i)'$ • the scatter matrix

Variance = eigevalue.

- The unit vector u that maximizes $u'Su$ is same as
  - the eigenvector with largest eigenvalue.
  - the eigenvector is called the first principal component

- Smmary of PCA
  - centering and normalization
  - finding a unit vector u, such that the variance of the projections along direction u is maximized
  - the variance is $\frac{1}{n} \cdot u'Su$
  - Computing the unit vector u which maximizes $u'Su$
    - eigenvector of S with the largest eigen value

1. centering and normalization
2. compute the scatter matrix
3. compute the eigenvectors & eigenvalues of the scatter matrix.

$$S = \sum_{i=1}^{n} x_i (x_i)'$$