Principal Component Analysis

Agenda

- Unsupervised Learning Approach
- Dimensionality Reduction
- Pre-requisite for fitting ML algorithms

Principal Component Analysis (PCA)

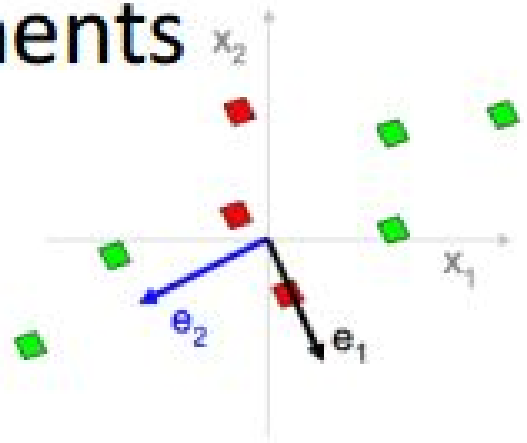
- Explain/summarize the underlying variance-covariance structure of a large set of variables through a few linear combinations of these variables

Applications

- Uses:
  - Data Visualization
  - Data Reduction
  - Data Classification
  - Trend Analysis
  - Factor Analysis
  - Noise Reduction

- Examples:
  - How many unique “sub-sets” are in the sample?
  - How are they similar / different?
  - What are the underlying factors that influence the samples?
  - Which time / temporal trends are (anti)correlated?
  - Which measurements are needed to differentiate?
  - How to best present what is “interesting”?
  - Which “sub-set” does this new sample rightfully belong?

Principal components



- Compute covariance matrix Σ
  - covariance of dimensions x_1 and x_2:
    - do x_1 and x_2 tend to increase together?
    - or does x_2 decrease as x_1 increases?
  - covariance: measure of variability

	x_1	x_2
x_1	2.0	0.8
x_2	0.8	0.6

$$\text{var}(x_2) = \frac{1}{N}\sum_{i=1}^{n}(x_{2,i} - \mu_2)^2$$

$$\text{cov}(x_1, x_2) = \frac{1}{N}\sum_{i=1}^{n}(x_{1,i} - \mu_1)(x_{2,i} - \mu_2)$$

- Find the basis of Σ
  - find vectors e_i which aren't turned by Σ
    - $\Sigma e_i = \lambda_i e_i$: eigenvalue / eigenvector
  - 1st PC: "longest" e_i (has largest λ_i), 2nd PC: next longest, …

		e_1	e_2
λ_1	0.26	x 0.4	-0.9
λ_2	2.42	y -0.9	-0.4