Lec 13

Wednesday, February 8, 2023 1:28 PM

Exploratory data analysis: PCA, Clustering

- you have already collected data from various methods we have discussed.
  - 100 material → [tensile strength, melting temp, bulk modulus, ...]
  - given material i, $x_i = [T, T_m, B, \dots]$

    $$X = N \text{ rows} \times d \text{ columns}$$

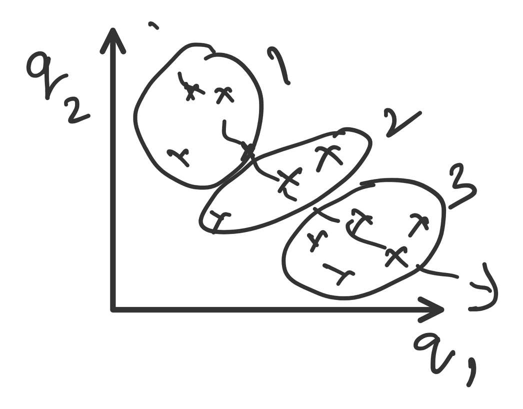
    ↑ "dimension"

  - d is typically high thus we need a way to visualize our summary statistics.
- mean array (vector) of our samples: $1 \times d$ vector

  covariance : $d \times d$

Dimensionality reduction:

- reduce $X_{N\times d} \longrightarrow Y_{N \times q}$ such that q is lower than d
- $q = 2$



- Principal component analysis (PCA)

  $$\boxed{f(\lambda) = \mu + V\lambda}$$
  $$= (1\times d) + (d \times q) \times (q \times 1)$$

  linear model parametrized by μ and V

  eg: $q = 2$ $[\lambda_0, \lambda_1] = \lambda$

  ↓

  $f(\lambda) = (1 \times d)$ vector

- $$\text{error} = \sum_{i=1}^{N} \| x_i - f(\lambda_i) \|^2$$
  $$= \sum_{i=1}^{N} \| x_i - \mu - V\lambda_i \|^2$$

  one solution that minimizes the above

  $$\hat{\mu} = \bar{x}_i$$
  $$\hat{\lambda}_i = V^T (x_i - \mu)$$
  $$= \sum_{i=1}^{N} \| x_i - \hat{\mu} - V V^T (x_i - \hat{\mu}) \|^2$$
  $$= \sum_{i=1}^{N} \| \tilde{x} - K \tilde{x} \|^2$$

  we are minimizing reconstruction error when projecting our mean centered samples by a matrix 'K'

- the projection by K maximizes the variance along each dimensions.

  $$X_{N\times d} \xrightarrow{K} Y_{N\times q}$$

  $$C_\lambda = \frac{1}{N} \Lambda \Lambda^T \quad \begin{cases} \Lambda = V^T (x - \mu) \\ \lambda_i = V^T (x_i - \mu) \end{cases}$$
  $$= \frac{1}{N} V^T \tilde{x} \tilde{x}^T V$$
  $$= \frac{1}{N} V^T C_{\tilde{x}} V$$

- spectral theorem $C_{\tilde{x}} = E^T D E$ ↳ diag(eigvals)

  $$C_\lambda = V^T E^T D E V$$
  $$= (EV)^T D (EV)$$

  our goal was to have 0 off diagonals for C_λ

  $$\boxed{V = E^{-1}} \Rightarrow C_\lambda = D \text{ diagonal matrix}$$

  $$\hat{\mu} = \tilde{x}$$
  $$\lambda_i = E(x_i - \hat{\mu})$$

  (E is an orthogonal matrix, $E^{-1} = E^T$, $V^T = E$)

- each dimension in the projected space has a variance given by eigenvalues in D which can be used to quantify explained variance

  ⇓

  find right set of dimension given a variance threshold

- generative model: (Asg 2)

  UV-vis spectroscopy ⟶ PCA ⟶ $[\lambda_0, \lambda_1, \lambda_2]$

  dim = wavelengths

  varying λ_0 ⟶ what are the features that are changing?

Clustering:

- group samples based on similarities (distance metric, similarity function)
- within group similarities to be maximum

  across group " minimum

- K-mean clustering: 1) assigning: grouping based on template

  convergence

  2) update: mean value of grouped points as template

  (when your templates does not change you stop iterating return the group labels)

- dim red : isomap, diffusion maps

  clustering: spectral clustering, density based clustering.

Elements of statistical learning, Ch 14