Unsupervised Learning
Principal Component Analysis

CMSC 422
Slides adapted from Prof. CARPUAT

Slides credit: Maria-Florina Balcan
Unsupervised Learning

• Discovering hidden structure in data

• What algorithms do we know for unsupervised learning?
  – K-Means Clustering

• Today: how can we learn better representations of our data points?
Dimensionality Reduction

• Goal: extract hidden lower-dimensional structure from high dimensional datasets

• Why?
  – To visualize data more easily
  – To remove noise in data
  – To lower resource requirements for storing/processing data
  – To improve classification/clustering
Low dimensional data embedded in high dimensional spaces

Examples of data points in D dimensional space that can be effectively represented in a d-dimensional subspace (d < D)
Principal Component Analysis

• Goal: Find a projection of the data onto directions that maximize variance of the original data set
  – Intuition: those are directions in which most information is encoded

• Definition: Principal Components are orthogonal directions that capture most of the variance in the data
PCA: finding principal components

- **1\textsuperscript{st} PC**
  - Projection of data points along 1\textsuperscript{st} PC discriminates data most along any one direction

- **2\textsuperscript{nd} PC**
  - next orthogonal direction of greatest variability

- And so on...
PCA: notation

• Data points
  – Represented by matrix X of size DxN
  – Let’s assume data is centered

• Principal components are d vectors: \( v_1, v_2, \ldots, v_d \)
  – \( v_i \cdot v_j = 0, i \neq j \) and \( v_i \cdot v_i = 1 \)

• The sample variance data projected on vector \( v \) is
  \[
  \frac{1}{n} \sum_{i=1}^{n} (v^T x_i)^2 = \frac{1}{n} v^T XX^T v
  \]
PCA formally

• Finding vector that maximizes sample variance of projected data:
  $$\arg\max_v v^T XX^T v \text{ such that } v^T v = 1$$

• A constrained optimization problem
  - Lagrangian folds constraint into objective:
    $$\arg\max_v v^T XX^T v - \lambda v^T v$$
  - Solutions are vectors $v$ such that $XX^T v = \lambda v$
    - i.e. eigenvectors of $XX^T$ (sample covariance matrix)
PCA formally

• The eigenvalue $\lambda$ denotes the amount of variability captured along dimension $\nu$
  – Sample variance of projection $\nu^TXX^T \nu = \lambda$

• If we rank eigenvalues from large to small
  – The 1$^{\text{st}}$ PC is the eigenvector of $XX^T$ associated with largest eigenvalue
  – The 2$^{\text{nd}}$ PC is the eigenvector of $XX^T$ associated with 2$^{\text{nd}}$ largest eigenvalue
  – ...

Alternative interpretation of PCA

• PCA finds vectors $v$ such that projection onto these vectors minimizes reconstruction error

$$\frac{1}{n} \sum_{i=1}^{n} \|x_i - (v^T x_i)v\|^2$$
Algorithm 36 \texttt{PCA}(D, K)

1. $\mu \leftarrow \text{MEAN}(X)$ \Comment{compute data mean for centering}
2. $D \leftarrow (X - \mu 1^\top)^\top (X - \mu 1^\top)$ \Comment{compute covariance, 1 is a vector of ones}
3. $\{\lambda_k, u_k\} \leftarrow \text{top } K \text{ eigenvalues/eigenvectors of } D$
4. \textbf{return} $(X - \mu 1) U$ \Comment{project data using $U$}
How to choose the hyperparameter $K$?

- i.e. the number of dimensions

- We can ignore the components of smaller significance
An example: Eigenfaces

Eigenfaces from 7562 images:

top left image is linear combination of rest.

Sirovich & Kirby (1987)
Turk & Pentland (1991)
PCA pros and cons

• Pros
  – Eigenvector method
  – No tuning of the parameters
  – No local optima

• Cons
  – Only based on covariance (2\textsuperscript{nd} order statistics)
  – Limited to linear projections
What you should know

• Principal Components Analysis
  – Goal: Find a projection of the data onto directions that maximize variance of the original data set
  – PCA optimization objectives and resulting algorithm
  – Why this is useful!