Unsupervised Learning
Principal Component Analysis

CMSC 422
MARINE CARPUAT
marine@cs.umd.edu

Illustrations 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
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\(^{\text{st}}\) PC**
  - Projection of data points along 1\(^{\text{st}}\) PC discriminates data most along any one direction

- **2\(^{\text{nd}}\) PC**
  - next orthogonal direction of greatest variability

- And so on...
PCA: notation

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

• Principal components are d vectors: \( v_1, v_2, \ldots, v_d \)
  \[ v_i \cdot v_j = 0, i \neq j \quad \text{and} \quad v_i \cdot v_i = 1 \]

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

• Finding vector that maximizes sample variance of projected data:
  \[ \underset{v}{\text{argmax}} \, v^T X^T X v \text{ such that } v^T v = 1 \]

• A constrained optimization problem
  ▪ Lagrangian folds constraint into objective:
    \[ \underset{v}{\text{argmax}} \, v^T X^T X v - \lambda (v^T v - 1) \]
  ▪ Solutions are vectors \( v \) such that \( X^T X v = \lambda v \)
    ▪ i.e. eigenvectors of \( X^T X \) (sample covariance matrix)
PCA formally

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

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

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

$$\frac{1}{n} \sum_{i=1}^{n} \|x_i - (v^T x_i)v\|^2$$
Resulting PCA algorithm

Algorithm 36 $\text{PCA}(D, K)$

1. $\mu \leftarrow \text{MEAN}(X)$ \hspace{1cm} // compute data mean for centering
2. $D \leftarrow \left( X - \mu \mathbf{1}^\top \right)^\top \left( X - \mu \mathbf{1}^\top \right)$ \hspace{1cm} // compute covariance, $\mathbf{1}$ is a vector of ones
3. $\{\lambda_k, u_k\} \leftarrow \text{top} \; K \; \text{eigenvalues/eigenvectors of} \; D$
4. $\text{return } (X - \mu \mathbf{1}) \; U$ \hspace{1cm} // 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!