PCA II

CMSC 422
SOHEIL FEIZI
sfeizi@cs.umd.edu
Today’s topics

• PCA

• 2\textsuperscript{nd} programming assignment posted
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
• Linear algebra review:
  – Matrix decomposition with eigenvectors and eigenvalues
Principal Component Analysis

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

• Definition: \textbf{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...
Examples of data points in D dimensional space that can be effectively represented in a d-dimensional subspace (d < D)
PCA: notation

• Data points
  – Represented by matrix $X$ of size $N \times D$
  – Let’s assume data is centered

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

• 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:

$$\arg\max \nu, \nu^T X^T X \nu \text{ such that } \nu^T \nu = 1$$

• A constrained optimization problem
  
  ▪ Lagrangian folds constraint into objective:
  $$\arg\max \nu, \nu^T X^T X \nu - \lambda(\nu^T \nu - 1)$$
  
  ▪ Solutions are vectors $\nu$ such that $X^T X \nu = \lambda \nu$
    ▪ i.e. eigenvectors of $X^T X$ (sample covariance matrix)
PCA formally

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

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

Alternative interpretation of PCA

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

**Algorithm 36 PCA(D, K)**

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