COMS 4771
Introduction to Machine Learning
Announcements

• Project presentations in one week!
  If you haven’t already, schedule a meeting with me this week

• HW4 posted, due next Tue.
Last time...

- Unsupervised Learning problems:
  Clustering and Dimensionality Reduction

- \( k \)-means

- Hierarchical Clustering

- Gaussian Mixture Models

- EM algorithm
Example: Handwritten digits

Handwritten digit data, but with no labels

What can we do?

- Suppose know that there are 10 groupings, can we find the groups?
- What if we don’t know there are 10 groups?
- How can we discover/explore other structure in such data?
Dimensionality Reduction

Data: \( \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d \)

Goal: find a ‘useful’ transformation \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^k \) that helps in the downstream prediction task.

Some previously seen useful transformations:

- **z-scoring** \( (x_1, \ldots, x_d) \mapsto \left( \frac{x_1 - \mu_1}{\sigma_1}, \ldots, \frac{x_d - \mu_d}{\sigma_d} \right) \)  
  Keeps same dimensionality but with better scaling

- **Kernel transformations.**  
  Higher dimensionality, making data linearly separable

**What are other desirable feature transformations?**

**How about lower dimensionality while keeping the relevant information?**
Principal Components Analysis (PCA)

Data: \( \vec{x}_1, \vec{x}_2, \ldots, \vec{x}_n \in \mathbb{R}^d \)

Goal: find the best **linear** transformation \( \phi : \mathbb{R}^d \rightarrow \mathbb{R}^k \) that best maintains reconstruction accuracy.

**Equivalently, minimize aggregate residual error**

Define: \( \Pi^k : \mathbb{R}^d \rightarrow \mathbb{R}^d \) \( k \)-dimensional orthogonal linear projector

\[
\text{minimize } \frac{1}{n} \sum_{i=1}^{n} \left\| \vec{x}_i - \Pi^k(\vec{x}_i) \right\|^2
\]

*How do we optimize this?*
A $k$ dimensional subspace can be represented by $\vec{q}_1, \ldots, \vec{q}_k \in \mathbb{R}^d$ orthonormal vectors.

The projection of any $\vec{x} \in \mathbb{R}^d$ in the $\text{span}(\vec{q}_1, \ldots, \vec{q}_k)$ is given by

$$
\left( \sum_{i=1}^{k} \vec{q}_i \vec{q}_i^T \right) \vec{x} = \sum_{i=1}^{k} (\vec{q}_i \cdot \vec{x}) \vec{q}_i
$$

To represent it in $\mathbb{R}^k$ (using basis $\vec{q}_1, \ldots, \vec{q}_k$) the coefficients simply are: $(\vec{q}_1 \cdot \vec{x}), \ldots, (\vec{q}_k \cdot \vec{x})$
PCA: $k = 1$ case

If projection dimension $k = 1$, then looking for a $q$ such that

$$
\text{minimize}_{\|q\|=1} \quad \frac{1}{n} \sum_{i=1}^{n} \left\| \vec{x}_i - (q \ q^T) \vec{x}_i \right\|^2
$$

$$
= \left( \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i^T \vec{x}_i \right) - q^T \left( \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i \vec{x}_i^T \right) q
$$

$$
\propto - q^T \left( \frac{1}{n} XX^T \right) q
$$

Equivalent formulation:

$$
\text{maximize}_{\|q\|=1} \quad q^T \left( \frac{1}{n} XX^T \right) q
$$

How to solve?
Recall for any matrix $M$, the $(\lambda, v)$ pairs of the fixed point equation

$$Mv = \lambda v$$

are the eigenvalue and the eigenvectors of $M$. ($v \neq 0$)

$$v^T M v = \lambda v^T v$$

$$\lambda = \frac{v^T M v}{v^T v} = \bar{v}^T M \bar{v} \quad \text{where} \quad \bar{v} = \frac{v}{\|v\|} \quad (\text{ie, unit length})$$

So,

$$\maximize_{\|q\|=1} \quad \bar{q}^T \left( \frac{1}{n} XX^T \right) \bar{q} \quad \text{Basically is the top eigenvector of matrix (1/n) \(XX^T\)!}$$
PCA: $k = 1$ case

$maximize_{\|q\|=1} \quad q^T \left( \frac{1}{n} XX^T \right) q$

Covariance of data (if mean = 0)

For any $q$ the quadratic form $q^T \left( \frac{1}{n} XX^T \right) q$ is the empirical variance of data in the direction $q$, i.e., of data $q^T \tilde{x}_1, \ldots, q^T \tilde{x}_n$

Therefore, the top eigenvector solution implies that the direction of maximum variance minimizes the residual error!

What about general $k$?
PCA: general \( k \) case

\[
\arg \min_{Q \in \mathbb{R}^{d \times k}} \frac{1}{n} \sum_{i=1}^{n} \left\| \vec{x}_i - QQ^T \vec{x}_i \right\|^2 = \arg \max_{Q \in \mathbb{R}^{d \times k}} \text{tr} \left( Q^T \left( \frac{1}{n} XX^T \right) Q \right)
\]

**Solution:** Basically is the top \( k \) eigenvectors of the matrix \( XX^T \)!

\[
\text{tr} \left( Q^T \left( \frac{1}{n} XX^T \right) Q \right) = \sum_{i=1}^{k} \text{empirical variance of } q_i^T x
\]

\( k \)-dimensional subspace preserving maximum amount of variance
PCA: Example Handwritten Digits

Images of handwritten 3s in $\mathbb{R}^{784}$

Mean

$q_1$
$\lambda_1 = 3.4 \cdot 10^5$

$q_2$
$\lambda_2 = 2.8 \cdot 10^5$

$q_3$
$\lambda_3 = 2.4 \cdot 10^5$

$q_4$
$\lambda_4 = 1.6 \cdot 10^5$

Any example:

$x = 3 + w_1 + w_2 + \ldots$

Data Reconstruction:

$x$
$k = 1$

$k = 10$

$k = 50$

$k = 200$

We can compress the each datapoint to just $k$ numbers!
Other Popular Dimension Reduction Methods

Multi-dimensional Scaling

Independent Component Analysis (ICA) (for blind source separation)

Non-negative matrix factorization (to create additive models)

Dictionary Learning

Random Projections

...  

All of them are linear methods
Non-Linear Dimensionality Reduction

Consider non-linear data

*Linear embedding*  
*non-linear embedding*
Basic optimization criterion:

Find an embedding that:

- Keeps neighboring points close
- Keeps far-off points far

Example variation 1:

*Distort neighboring distances by at most $(1 \pm \varepsilon)$ factor, while maximizing non-neighbor distances.*

Example variation 2:

*Compute *geodesic* (local hop) distances, and find an embedding that best preserves geodesics.*
Non-linear embedding: Example
Popular Non-Linear Methods

Locally Linear Embedding (LLE)

Isometric Mapping (IsoMap)

Laplacian Eigenmaps (LE)

Local Tangent Space Alignment (LTSA)

Maximum Variance Unfolding (MVU)

...
Probabilistic Reasoning via Graphical Models
A probabilistic model where a graph represents the conditional dependence structure among the variables.

Example:
Four variables of interest – cloudiness, raining, sprinkler, grass_wet

Possible Causality Structure:

Cloudy (C) → Rainy (R) → GrassWet (G) → Sprinkler (S)

Inference questions:
• What is the probability it rained given the grass is wet?
• What is the chance that the sprinkler was off given grass is wet and it is not cloudy?

Learning questions:
• What is the most likely GM structure and connection weights that models the data?
Graphical Models: Representation

There are two kinds of Graphical Models:

Directed models – Bayesian Networks

Undirected models – Markov Random Fields (MRFs)
Bayesian Networks

What is the joint probability for these variables?

\[ P(C, S, R, G) \]

\[ = P(C)P(R|C)P(S|R, C)P(G|S, R, C) \quad \text{Chain rule} \]

\[ = P(C)P(R|C)P(S|C)P(G|S, R) \quad \text{due to the (in)dependencies asserted by the parent-child relationships} \]

**In general:**

\[ P(X_1, \ldots, X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parent}(X_i)) \]

*That is: a variable is independent of its ancestors given the parents.*
These conditional probability tables (CPT) are enough to **completely** specify the joint distribution!
Bayesian Networks: Inference

\[ P(C, S, R, G) = P(C)P(R|C)P(S|C)P(G|S, R) \]

Q: What is the probability of sprinkler being on given the grass is wet?

\[ P(S = 1|G = 1) = \frac{P(S = 1, G = 1)}{P(G = 1)} = \frac{0.2781}{0.6471} = 0.430 \]

\[ P(G = 1) = \sum_{c,s,r} P(C = c, S = s, R = r, G = 1) \]

\[ = \sum_{c,s,r} P(C = c)P(R = r|C = c)P(S = s|C = c)P(G = 1|S = s, R = r) \]

\[ = 0.6471 \]

\[ P(S = 1, G = 1) = \sum_{c,r} P(C = c, S = 1, R = r, G = 1) = \cdots = 0.2781 \]
Bayesian Networks: Learning Parameters

\[ P(C, S, R, G) = P(C)P(R|C)P(S|C)P(G|S, R) \]

Learning the parameters knowing the structure

*ie, estimate the CPTs from observations*

Simply do the likelihood estimates (ie, counts)

\[ \hat{P}_{ML}(G = g|S = s, R = r) = \frac{\#(G = g, S = s, R = r)}{\#(S = s, R = r)} \]

etc ...

*Issue: assigns zero prob. for unseen combinations in data. How to fix that?*
Bayesian Networks: Learning Structure

\[ P(C, S, R, G) = P(C)P(R|C)P(S|R, C)P(G|S, R, C) \]

Learning the unknown structure between the variables

General

- Test of conditional independencies in data
- Grow-Shrink Markov Blanket algorithm

Assumed structure:

- Tree structure: Chow-Liu algorithm
- Small cliques: variations on Chow-Liu

NP-hard to find the optimal structure
Markov Random Fields (MRFs)

Graphical models with undirected connections

\[ P(X_1, \ldots, X_d) = \frac{1}{Z} \prod_{C \in \text{max-cliques}(G)} \phi_C(X_C) \]

normalizer (so things integrate to 1), aka the partition function

Clique potentials, typically the relative frequency of variable co-occurrence in a clique

Example: five variable graph

\[ P(X_1, \ldots, X_5) \propto \phi_1((X_1, X_2, X_3)) \phi_2((X_2, X_3, X_4)) \phi_3((X_3, X_5)) \]

What are the max-cliques?
What are the (conditional) independencies asserted by the following directed model?

What is the equivalent undirected model?
What We Learned...

- Dimensionality Reduction
  - Linear vs non-linear Dimensionality Reduction
- Principal Component Analysis
- Graphical Models
- Bayesian Networks
- Doing inference and learning on Bayes Nets
Questions?
Next time...

Continue Graphical Models

Markov Models

Hidden Markov Models
Announcements

• Project presentations in one week!
  If you haven’t already, schedule a meeting with me this week

• HW4 posted, due next Tue.