Lecture 24:

- Autoencoders
- ICA
Last time... Dimensionality Reduction

- Clustering
  - One way to summarize a complex real-valued data point with a single categorical variable

- Dimensionality reduction
  - Another way to simplify complex high-dimensional data
  - Summarize data with a lower dimensional real valued vector

- Given data points in $d$ dimensions
- Convert them to data points in $r<d$ dims
- With minimal loss of information
Last time... Principal Component Analysis

- **PCA Vectors** originate from the center of mass.

- Principal component #1: points in the direction of the **largest variance**.

- Each subsequent principal component
  - is **orthogonal** to the previous ones, and
  - points in the directions of the **largest variance of the residual subspace**
Last time... PCA Applications

Face Recognition

Image Compression

\[ x \xrightarrow{U} x' \]

Noise Filtering

16 most important eigenvectors
Today

• PCA shortcomings
• Autoencoders
• ICA
PCA Shortcomings
Problems Data Set for PCA

- PCA doesn’t know labels!
PCA vs. Fisher Linear Discriminant

**Principal Component Analysis**
- higher variance
- bad for discriminability

**Fisher Linear Discriminant**
- smaller variance
- good discriminability

---

slide by Javier Hernandez Rivera
Problematic Data Set for PCA

- PCA cannot capture NON-LINEAR structure!
PCA Conclusions

• PCA
  - Finds orthonormal basis for data
  - Sorts dimensions in order of “importance”
  - Discard low significance dimensions

• Uses:
  - Get compact description
  - Ignore noise
  - Improve classification (hopefully)

• Not magic:
  - Doesn’t know class labels
  - Can only capture linear variations

• One of many tricks to reduce dimensionality!
Autoencoders
Relation to Neural Networks

- PCA is closely related to a particular form of neural network
- An autoencoder is a neural network whose outputs are its own inputs
- The goal is to minimize reconstruction error
Auto encoders

- Define

\[ z = f(Wx); \quad \hat{x} = g(Vz) \]
Auto encoders

• Define

\[ z = f(Wx); \quad \hat{x} = g(Vz) \]

• Goal:

\[
\min_{W,V} \frac{1}{2N} \sum_{n=1}^{N} ||x^{(n)} - \hat{x}^{(n)}||^2
\]

In other words, the optimal solution is PCA.

Urtasun, Zemel, Fidler (UofT)

CSC 411: 14-PCA & Autoencoders

March 14, 2016 16 / 18
Auto encoders

- Define

\[ z = f(Wx); \quad \hat{x} = g(Vz) \]

- Goal:

\[
\min_{w,v} \quad \frac{1}{2N} \sum_{n=1}^{N} \| x^{(n)} - \hat{x}^{(n)} \|^2
\]

- If \( g \) and \( f \) are linear

\[
\min_{w,v} \quad \frac{1}{2N} \sum_{n=1}^{N} \| x^{(n)} - VWx^{(n)} \|^2
\]

In other words, the optimal solution is PCA.

Urtasun, Zemel, Fidler (UofT)

CSC 411: 14-PCA & Autoencoders

March 14, 2016 16 / 18
Auto encoders

• Define

\[ z = f(Wx); \quad \hat{x} = g(Vz) \]

• Goal:

\[
\min_{w,v} \quad \frac{1}{2N} \sum_{n=1}^{N} ||x^{(n)} - \hat{x}^{(n)}||^2
\]

• If \( g \) and \( f \) are linear

\[
\min_{w,v} \quad \frac{1}{2N} \sum_{n=1}^{N} ||x^{(n)} - VWx^{(n)}||^2
\]

• In other words, the optimal solution is PCA
Auto encoders: Nonlinear PCA

• What if $g()$ is not linear?
• Then we are basically doing nonlinear PCA
• Some subtleties but in general this is an accurate description
Comparing Reconstructions

- Real data
- 30-d deep autoencoder
- 30-d logistic PCA
- 30-d PCA
Independent Component Analysis (ICA)
A Serious Limitation of PCA

- Recall that PCA looks at the covariance matrix only. What if the data is not well described by the covariance matrix?

- The only distribution which is uniquely specified by its covariance (with the subtracted mean) is the Gaussian distribution. Distributions which deviate from the Gaussian are poorly described by their covariances.
Faithful vs Meaningful Representations

• Even with non-Gaussian data, variance maximization leads to the most faithful representation in a reconstruction error sense (recall that we trained our autoencoder network using a mean-square error in an input reconstruction layer).

• The mean-square error measure implicitly assumes Gaussianity, since it penalizes datapoints close to the mean less that those that are far away.

• But it does not in general lead to the most meaningful representation.

• We need to perform gradient descent in some function other than the reconstruction error.
A Criterion Stronger than Decorrelation

- The way to circumvent these problems is to look for components which are statistically independent, rather than just uncorrelated.

- For statistical independence, we require that
  \[ p(\xi_1, \xi_2, \ldots, \xi_N) = \prod_{i=1}^{N} p(\xi_i) \]

- For uncorrelatedness, all we required was that
  \[ \langle \xi_i\xi_j \rangle - \langle \xi_i \rangle \langle \xi_j \rangle = 0 , \quad i \neq j \]

- Independence is a stronger requirement; under independence,
  \[ \langle g_1(\xi_i)g_2(\xi_j) \rangle - \langle g_1(\xi_i) \rangle \langle g_2(\xi_j) \rangle = 0 , \quad i \neq j \]
  for any functions \( g_1 \) and \( g_2 \).
Independent Component Analysis (ICA)

- Like PCA, except that we’re looking for a transformation subject to the stronger requirement of independence, rather than uncorrelatedness.

- In general, no analytic solution (like eigenvalue decomposition for PCA) exists, so ICA is implemented using neural network models.

- To do this, we need an architecture and an objective function to descend/climb in.

- Leads to $N$ independent (or as independent as possible) components in $N$-dimensional space; they need not be orthogonal.

- When are independent components identical to uncorrelated (principal) components? When the generative distribution is uniquely determined by its first and second moments. This is true of only the Gaussian distribution.
Neural Network for ICA

- Single layer network:

- Patterns \( \{\xi\} \) are fed into the input layer.
- Inputs multiplied by weights in matrix \( \mathbf{W} \).
- Output logistic (vector notation here):

\[
\bar{y} = \frac{1}{1 + e^{\mathbf{W}^T \bar{\xi}}}
\]
Objective Function for ICA

- Want to ensure that the outputs $y_i$ are maximally independent.
- This is identical to requiring that the mutual information be small. Or alternately that the joint entropy be large.

$$H(p) = \text{entropy of distribution } p \text{ of first neuron’s output}$$

$$H(p|q) = \text{conditional entropy}$$

$$I(p; q) = H(p) - H(q|p) = H(q) - H(p|q) = \text{mutual information}$$

- Gradient ascent in this objective function is called infomax (we’re trying to maximize the enclosed area representing information quantities).
Blind Source Separation (BSS)

- The most famous application of ICA.

- Have $K$ sources $\{s_k[t]\}$, and $K$ signals $\{x_k[t]\}$. Both $\{s_k[t]\}$ and $\{x_k[t]\}$ are time series ($t$ is a discrete time index).

- Each signal is a linear mixture of the sources

$$x_k[t] = A_s k[t] + n_k[t]$$

where $n_k[t]$ is the noise contribution in the kth signal $x_k[t]$, and $A$ is a mixture matrix.

- The problem: given $x_k[n]$, determine $A$ and $s_k[n]$.
The Cocktail Party

Sources

Mixing

Observation

ICA Estimation

\[ s(t) \]

\[ A \in \mathbb{R}^{M \times M} \]

\[ x(t) = As(t) \]

\[ y(t) = Wx(t) \]
Demo: The Cocktail Party

- Frequency domain ICA (1995)

Input mix:  

Extracted speech:  

http://paris.cs.illinois.edu/demos/index.html