Unsupervised Learning
PCA

Marek Petrik

3/2/2017
Learning Methods

1. **Supervised Learning**: Learning a function $f$:

   \[ Y = f(X) + \epsilon \]

   1.1 Regression
   1.2 Classification

2. **Unsupervised learning**: Discover interesting properties of data (no labels)

   \[ X_1, X_2, \ldots \]

   2.1 Dimensionality reduction or embedding
   2.2 Clustering
Principal Components Analysis

- Reduce dimensionality
- Start with features $X_1 \ldots X_n$
- Construct *fewer* features $Z_1 \ldots Z_M$

\[
Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p
\]

- Weights are usually normalized (using $\ell_2$ norm)

\[
\sum_{j=1}^{p} \phi_{j1}^2 = 1
\]

- Data has greatest variance along $Z_1$
1st Principal Component

- **1st Principal Component**: Direction with the largest variance

\[ Z_1 = 0.839 \times (\text{pop} - \bar{\text{pop}}) + 0.544 \times (\text{ad} - \bar{\text{ad}}) \]
1st Principal Component

1st Principal Component: Direction with the largest variance

\[ Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}}) \]

Is this linear?
1st Principal Component

1st Principal Component: Direction with the largest variance

\[ Z_1 = 0.839 \times (\text{pop} - \overline{\text{pop}}) + 0.544 \times (\text{ad} - \overline{\text{ad}}) \]

Is this linear? Yes, after *mean centering*. 
green line: 1st principal component, minimize distances to all points
green line: 1st principal component, minimize distances to all points

Is this the same as linear regression?
green line: 1st principal component, minimize distances to all points

Is this the same as linear regression? **No**, like *total least squares*. 
2nd Principal Component

- **2nd Principal Component**: Orthogonal to 1st component, largest variance

\[ Z_2 = 0.544 \times (\text{pop} - \bar{\text{pop}}) - 0.839 \times (\text{ad} - \bar{\text{ad}}) \]
1st Principal Component

Population

Ad Spending

2nd Principal Component

Population

Ad Spending
Solving PCA

\[
\min_{\phi_1, \ldots, \phi_p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\}
\]

subject to \( \sum_{j=1}^{p} \phi_{j1}^2 = 1 \)

Solve using eigenvalue decomposition
Interpretation of 1st Principal Component

1. Direction with the largest variance

2. Line with smallest distance to all points
PCA Example
PCA Technicalities

1. Features should be centered = zero mean

2. Scale of features matters

3. The direction (sign) of principal vectors is not unique

4. Proportion of Variance Explained: variance along the dimension / total variance

5. How many principal vectors?
1. Features should be **centered** = zero mean

2. **Scale** of features matters

3. The direction (sign) of principal vectors is not unique

4. **Proportion of Variance Explained**: variance along the dimension / total variance

5. How many principal vectors? It depends …
Partial Least Squares

- Supervised version of PCR
Problem With High Dimensions

- Computational complexity
- Overfitting is a problem
Overfitting with Many Variables

- $R^2$ vs. Number of Variables
- Training MSE vs. Number of Variables
- Test MSE vs. Number of Variables
Examples

1. Simple PCA: R notebook

2. MNIST PCA: https://colah.github.io/posts/2014-10-Visualizing-MNIST/