Lecture Slides for

Machine Learning

2nd Edition

CHAPTER 6:

Dimensionality Reduction

ETHEM ALPAYDIN

© The MIT Press, 2010

alpaydin@boun.edu.tr

http://www.cmpe.boun.edu.tr/~ethem/i2ml2e
Why Reduce Dimensionality?

- Reduces time complexity: Less computation
- Reduces space complexity: Less parameters
- Saves the cost of observing the feature
- Simpler models are more robust on small datasets
- More interpretable; simpler explanation
- Data visualization (structure, groups, outliers, etc) if plotted in 2 or 3 dimensions
Feature Selection vs Extraction

- **Feature selection:** Choosing $k<d$ important features, ignoring the remaining $d - k$
  
  Subset selection algorithms

- **Feature extraction:** Project the original $x_i$, $i = 1,...,d$ dimensions to new $k<d$ dimensions, $z_j$, $j = 1,...,k$

Principal components analysis (PCA), linear discriminant analysis (LDA), factor analysis (FA)
Subset Selection

- There are $2^d$ subsets of $d$ features
- Forward search: Add the best feature at each step
  - Set of features $F$ initially $\emptyset$.
  - At each iteration, find the best new feature
    \[ j = \arg\min_i E(F \cup x_i) \]
  - Add $x_j$ to $F$ if $E(F \cup x_j) < E(F)$

- Hill-climbing $O(d^2)$ algorithm
- Backward search: Start with all features and remove one at a time, if possible.
- Floating search (Add $k$, remove $l$)
Principal Components Analysis (PCA)

- Find a low-dimensional space such that when \( x \) is projected there, information loss is minimized.
- The projection of \( x \) on the direction of \( w \) is: \( z = w^T x \)
- Find \( w \) such that \( \text{Var}(z) \) is maximized

\[
\text{Var}(z) = \text{Var}(w^T x) = \mathbb{E}[(w^T x - w^T \mu)^2]
\]
\[
= \mathbb{E}[(w^T x - w^T \mu)(w^T x - w^T \mu)]
\]
\[
= \mathbb{E}[w^T(x - \mu)(x - \mu)^T w]
\]
\[
= w^T \mathbb{E}[(x - \mu)(x - \mu)^T]w = w^T \Sigma w
\]

where \( \text{Var}(x) = \mathbb{E}[(x - \mu)(x - \mu)^T] = \Sigma \)
Maximize \( \text{Var}(z) \) subject to \( ||w|| = 1 \)

\[
\max_{w_1} w_1^T \Sigma w_1 - \alpha (w_1^T w_1 - 1)
\]

\( \Sigma w_1 = \alpha w_1 \) that is, \( w_1 \) is an eigenvector of \( \Sigma \)

Choose the one with the largest eigenvalue for \( \text{Var}(z) \) to be max

Second principal component: Max \( \text{Var}(z_2) \), s.t., \( ||w_2|| = 1 \) and orthogonal to \( w_1 \)

\[
\max_{w_2} w_2^T \Sigma w_2 - \alpha (w_2^T w_2 - 1) - \beta (w_2^T w_1 - 0)
\]

\( \Sigma w_2 = \alpha w_2 \) that is, \( w_2 \) is another eigenvector of \( \Sigma \)

and so on.
What PCA does

\[ z = W^T(x - m) \]

where the columns of \( W \) are the eigenvectors of \( \Sigma \), and \( m \) is sample mean.

Centers the data at the origin and rotates the axes.
How to choose k?

• Proportion of Variance (PoV) explained

\[
\frac{\lambda_1 + \lambda_2 + \lambda_k}{\lambda_1 + \lambda_2 + \lambda_k + \lambda_d}
\]

when \( \lambda_i \) are sorted in descending order

• Typically, stop at PoV > 0.9

• Scree graph plots of PoV vs k, stop at “elbow”
Factor Analysis

- Find a small number of factors $\mathbf{z}$, which when combined generate $\mathbf{x}$:

$$x_i - \mu_i = v_{i1}z_1 + v_{i2}z_2 + ... + v_{ik}z_k + \varepsilon_i$$

where $z_j, j = 1, ..., k$ are the latent factors with

$$E[ z_j ]=0, \ Var(z_j)=1, \ Cov(z_i, z_j)=0, \ i \neq j,$$

$\varepsilon_i$ are the noise sources

$$E[ \varepsilon_i ]= \psi_i, \ Cov(\varepsilon_i, \varepsilon_j) =0, \ i \neq j, \ Cov(\varepsilon_i, z_j) =0,$$

and $v_{ij}$ are the factor loadings
PCA vs FA

- **PCA** From \( x \) to \( z \) \[ z = W^T(x - \mu) \]
- **FA** From \( z \) to \( x \) \[ x - \mu = Vz + \varepsilon \]
Factor Analysis

- In FA, factors $z_j$ are stretched, rotated and translated to generate $x$
Multidimensional Scaling

- Given pairwise distances between $N$ points, $d_{ij}$, $i,j = 1, \ldots, N$
  place on a low-dim map s.t. distances are preserved.

- $z = g(x | \theta)$ Find $\theta$ that min Sammon stress

$$E(\theta | X) = \sum_{r,s} \frac{\left( \frac{\left| z^r - z^s \right| - \left| x^r - x^s \right|}{\left| x^r - x^s \right|^2} \right)^3}{\frac{\left| x^r - x^s \right|^2}{\left| x^r - x^s \right|^2}}$$

$$= \sum_{r,s} \frac{\left( \left| g(x^r | \theta) - g(x^s | \theta) \right| - \left| x^r - x^s \right| \right)^3}{\left| x^r - x^s \right|^2}$$
Map of Europe by MDS

Map from CIA – The World Factbook: http://www.cia.gov/
Linear Discriminant Analysis

• Find a low-dimensional space such that when $\mathbf{x}$ is projected, classes are well-separated.

• Find $\mathbf{w}$ that maximizes

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2}$$

$$m_1 = \frac{\sum_t \mathbf{w}^T \mathbf{x}^t r^t}{\sum_t r^t}$$

$$s_1^2 = \sum_t \left(\mathbf{w}^T \mathbf{x}^t - m_1\right)^2 r^t$$
Between-class scatter:

\[(m_1 - m_2)^2 = (w^T m_1 - w^T m_2)^2\]
\[= w^T (m_1 - m_2)(m_1 - m_2)^T w\]
\[= w^T S_B w \text{ where } S_B = (m_1 - m_2)(m_1 - m_2)^T\]

Within-class scatter:

\[s_1^2 = \sum_t (w^T x^t - m_1)^2 r^t\]
\[= \sum_t w^T (x^t - m_1)(x^t - m_1)^T w r^t = w^T S_1 w\]

where \(S_1 = \sum_t (x^t - m_1)(x^t - m_1)^T r^t\)

\[s_1^2 + s_1^2 = w^T S_w w \text{ where } S_w = S_1 + S_2\]
Fisher’s Linear Discriminant

• Find $\mathbf{w}$ that max

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} = \frac{\left| \mathbf{w}^T (\mathbf{m}_1 - \mathbf{m}_2) \right|^2}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

• LDA soln:

$$\mathbf{w} = c \times \mathbf{S}_W^{-1} (\mathbf{m}_1 - \mathbf{m}_2)$$

• Parametric soln:

$$\mathbf{w} = \Sigma^{-1} (\bar{x}_1 - \bar{x}_2)$$

when $p(\mathbf{x} | C_i) \sim \mathcal{N}(\bar{x}_i, \Sigma)$
K>2 Classes

- Within-class scatter:
  \[ S_W = \sum_{i=1}^{K} S_i \]
  \[ S_i = \sum_{t} r_i^t (x_t^t - m_i)(x_t^t - m_i)^\top \]

- Between-class scatter:
  \[ S_B = \sum_{i=1}^{K} N_i (m_i - m)(m_i - m)^\top \]
  \[ m = \frac{1}{K} \sum_{i=1}^{K} m_i \]

- Find \( W \) that max
  \[ J(W) = \frac{\left| W^T S_B W \right|}{\left| W^T S_W W \right|} \]
  The largest eigenvectors of \( S_W^{-1} S_B \)
  Maximum rank of \( K-1 \)
Isomap

- Geodesic distance is the distance along the manifold that the data lies in, as opposed to the Euclidean distance in the input space.
Isomap

• Instances $r$ and $s$ are connected in the graph if
  \[ \| x_r - x_s \| < \varepsilon \text{ or if } x_s \text{ is one of the } k \text{ neighbors of } x_r \]

  The edge length is \[ \| x_r - x_s \| \]

• For two nodes $r$ and $s$ not connected, the distance is equal to the shortest path between them

• Once the $N \times N$ distance matrix is thus formed, use MDS to find a lower-dimensional mapping
Locally Linear Embedding

1. Given $x^r$ find its neighbors $x^s_{(r)}$

2. Find $W_{rs}$ that minimize

$$E(W | X) = \sum_r \left\| x^r - \sum_s W_{rs} x^s_{(r)} \right\|^2$$

3. Find the new coordinates $z^r$ that minimize

$$E(z | W) = \sum_r \left\| z^r - \sum_s W_{rs} z^s_{(r)} \right\|^2$$
LLE on Optdigits