EE6550 Machine Learning

Lecture Nine – Dimensionality Reduction

Chung-Chin Lu

Department of Electrical Engineering

National Tsing Hua University

May 8, 2017

Motivations

- Computational: to compress the initial data as a preprocessing step to speed up subsequent operations on the data.
- Visualization: to visualize the data for exploratory analysis by mapping the input data into two- or three-dimensional spaces.
- Feature extraction: to hopefully generate a smaller and more effective or useful set of features

Formulation of Dimensionality Reduction

- $\mathscr{I}$: the input space of all items, associated with a probability space $(\mathscr{I}, \mathcal{F}, D)$.
- $\Phi : \mathscr{I} \to \mathbb{R}^N$: a feature mapping from the input space $\mathscr{I}$ to the N-dimensional feature space $\mathbb{R}^N$.
- $S = (\omega_1, \ldots, \omega_m)$: a sample of m items, drawn i.i.d. from the input space according to D.
- $\mathbf{X} = [\Phi(\omega_1), \ldots, \Phi(\omega_m)]$: the $N \times m$ data matrix associated with the sample S.
- Problem: to find, for $k \ll N$, a k-dimensional representation of the data, $\mathbf{Y} \in \mathbb{R}^{k \times m}$, that is in some way faithful to the original representation $\mathbf{X}$.

The Contents of This Lecture

- Principal component analysis
- Kernel principal component analysis
- Johnson-Lindenstrauss lemma

Formulation of Principal Component Analysis

- k : a fixed integer in $[1, N]$.
- $\mathcal{P}_k$: the family of all rank-k $N \times N$ orthogonal projection matrices.
- $\mathbf{X} = [\Phi(\omega_1), \ldots, \Phi(\omega_m)]$: an $N \times m$ data matrix associated with a random sample S of size m.
- Problem : to project the input data $\mathbf{X} = [\Phi(\omega_1), \ldots, \Phi(\omega_m)]$ in the N-dimensional feature space $\mathbb{R}^N$ onto the data in a k-dimensional linear subspace of $\mathbb{R}^N$ that minimizes the sum of the squared L_2-distances between the original data and the projected data,

$$\min_{\mathbf{P} \in \mathcal{P}_k} \|\mathbf{PX} - \mathbf{X}\|_F^2 = \min_{\mathbf{P} \in \mathcal{P}_k} \sum_{i=1}^{m} \|\mathbf{P}\Phi(\omega_i) - \Phi(\omega_i)\|^2.$$

- $\|\mathbf{A}\|_F \triangleq \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2}$ is called the Frobenius norm of an $n \times m$ matrix $\mathbf{A}$. Note that $\|\mathbf{A}\|_F^2 = \mathrm{tr}(\mathbf{A}\mathbf{A}^T) = \mathrm{tr}(\mathbf{A}^T\mathbf{A})$.
- Any $\mathbf{P}^* \in \arg\min_{\mathbf{P}\in\mathcal{P}_k} \|\mathbf{P}\mathbf{X} - \mathbf{X}\|_F^2$ is called a PCA solution of the data matrix $\mathbf{X}$.
- The k-dimensional column space of a PCA solution $\mathbf{P}^*$ has an orthonormal basis $\mathbf{u}_1^*, \ldots, \mathbf{u}_k^*$ so that

$$\mathbf{P}^*\Phi(\omega_i) = \sum_{j=1}^{k} y_{ji}\mathbf{u}_j^* = [\mathbf{u}_1^* \ldots \mathbf{u}_k^*] \begin{bmatrix} y_{1i} \\ \vdots \\ y_{ki} \end{bmatrix} = \mathbf{U}_k^*\mathbf{y}_i \;\forall\; i \in [1, m],$$

  where $\mathbf{U}_k^* = [\mathbf{u}_1^*, \ldots, \mathbf{u}_k^*]$ and $\mathbf{y}_i = [y_{1i}, \ldots, y_{ki}]^T$.
- Thus $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_m]$ is an optimal k-dimensional representation of the data $\mathbf{X}$, whose reconstruction $\mathbf{U}_k^*\mathbf{Y} = \mathbf{P}^*\mathbf{X}$ of $\mathbf{X}$ minimizes the squared error.

Cyclic Permutation Property of the Trace Function

For any $m \times n$ matrix $\mathbf{A}$ and any $n \times m$ matrix $\mathbf{B}$, we have

$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA}).$$

Proof.
$\mathrm{tr}(\mathbf{AB}) = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ji} = \sum_{j=1}^{n} \sum_{i=1}^{m} B_{ji} A_{ij} = \mathrm{tr}(\mathbf{BA}).$ □

- If $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_{n-1}, \mathbf{A}_n$ be $m_1 \times m_2$, $m_2 \times m_3$, …, $m_{n-1} \times m_n$, $m_n \times m_1$ matrices, then

$$\mathrm{tr}(\mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_{n-1} \mathbf{A}_n) = \mathrm{tr}(\mathbf{A}_n \mathbf{A}_1 \mathbf{A}_2 \cdots \mathbf{A}_{n-1}).$$

Orthogonal Projection Matrices

An $N \times N$ matrix $\mathbf{P}$ is a nonzero orthogonal projection matrix if and only if there exists an $N \times k$ matrix $\mathbf{U}_k$, $k \in [1, N]$, with $\mathbf{U}_k^T \mathbf{U}_k = I_{k \times k}$ such that $\mathbf{P} = \mathbf{U}_k \mathbf{U}_k^T$. Furthermore, the column space of $\mathbf{P}$ is the same as the column space of $\mathbf{U}_k$.

Proof. "$\Rightarrow$" Let the rank of $\mathbf{P}$ is k. Since $\mathbf{P}$ is nonzero, $k \in [1, N]$. Since $\mathbf{P}^2 = \mathbf{P}$, the eigenvalues of $\mathbf{P}$ can be 0 or 1 only. Since $\mathbf{P}$ is symmetric, $\mathbf{P}$ has a spectral decomposition

$$\mathbf{P} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$$

where $\mathbf{U}$ is an $N \times N$ orthogonal matrix and $\mathbf{\Lambda}$ is a diagonal matrix with 1's in the first k diagonal entries in the diagonal and 0's in the rest diagonal entries. Thus we have

$$\mathbf{P} = \mathbf{U}_k \mathbf{U}_k^T,$$

where $\mathbf{U}_k$ consists of the first k columns of $\mathbf{U}$. Since $\mathbf{U}$ is orthogonal, we have $\mathbf{U}_k^T\mathbf{U}_k = I_{k\times k}$. "$\Leftarrow$" We check that

$$\mathbf{P}^T = (\mathbf{U}_k\mathbf{U}_k^T)^T = \mathbf{U}_k\mathbf{U}_k^T = \mathbf{P}$$

and

$$\mathbf{P}^2 = \mathbf{U}_k(\mathbf{U}_k^T\mathbf{U}_k)\mathbf{U}_k^T = \mathbf{U}_k I_{k\times k}\mathbf{U}_k^T = \mathbf{P}.$$

∞

Since $k > 0$, $\mathbf{P} = \mathbf{U}_k\mathbf{U}_k^T$ is a nonzero orthogonal projection matrix. The column space of $\mathbf{P}$ is

$$\{\mathbf{P}\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^N\} = \{\mathbf{U}_k\mathbf{U}_k^T\mathbf{x} \mid \mathbf{x} \in \mathbb{R}^N\} = \{\mathbf{U}_k\mathbf{y} \mid \mathbf{y} \in \mathbb{R}^k\},$$

which is the column space of $\mathbf{U}_k$, since the rank of $\mathbf{U}_k^T$ is k. □

PCA Solution

Theorem 12.1: Let

- $\mathbf{X} = [\Phi(\omega_1), \ldots, \Phi(\omega_m)]$: the $N \times m$ data matrix associated with a random sample $S = (\omega_1, \ldots, \omega_m)$.
- $\mathbf{C} \triangleq \frac{1}{m}\mathbf{X}\mathbf{X}^T$: the covariance matrix of the data matrix $\mathbf{X}$.
- $\mathbf{U}_k^* = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$: an $N \times k$ matrix consisting of orthonormal eigenvectors corresponding to the k largest eigenvalues of the covariance matrix C.

A PCA solution is

$$\mathbf{P}^* = \mathbf{U}_k^* \mathbf{U}_k^{*T}$$

and the associated k-dimensional representation of $\mathbf{X}$ is given by

$$\mathbf{Y} = \mathbf{U}_k^{*T} \mathbf{X}.$$

Proof. Since

$$\begin{aligned} \|\mathbf{PX}-\mathbf{X}\|_F^2 &= \mathrm{tr}((\mathbf{PX}-\mathbf{X})^T(\mathbf{PX}-\mathbf{X})) \\ &= \mathrm{tr}(\mathbf{X}^T\mathbf{P}^2\mathbf{X}-2\mathbf{X}^T\mathbf{PX}+\mathbf{X}^T\mathbf{X}) \\ &= -\mathrm{tr}(\mathbf{X}^T\mathbf{PX})+\mathrm{tr}(\mathbf{X}^T\mathbf{X}) \end{aligned}$$

for any $N \times N$ orthogonal projection matrix $\mathbf{P}$ and $\mathrm{tr}(\mathbf{X}^T\mathbf{X})$ is independent of $\mathbf{P}$, we have

$$\arg\min_{\mathbf{P}\in\mathcal{P}_k} \|\mathbf{PX}-\mathbf{X}\|_F^2 = \arg\max_{\mathbf{P}\in\mathcal{P}_k} \mathrm{tr}(\mathbf{X}^T\mathbf{PX}).$$

For any $\mathbf{P} \in \mathcal{P}_k$, there is an $N \times k$ matrix $\mathbf{U}_k$ with $\mathbf{U}_k^T\mathbf{U}_k = I_{k\times k}$ such that $\mathbf{P} = \mathbf{U}_k\mathbf{U}_k^T$ and then

$$\mathrm{tr}(\mathbf{X}^T\mathbf{PX}) = \mathrm{tr}(\mathbf{X}^T\mathbf{U}_k\mathbf{U}_k^T\mathbf{X}) = \mathrm{tr}(\mathbf{U}_k^T\mathbf{X}\mathbf{X}^T\mathbf{U}_k) = \sum_{i=1}^{k} \mathbf{u}_i^T(\mathbf{X}\mathbf{X}^T)\mathbf{u}_i$$

by the cyclic permutation property of the trace function, where $\mathbf{u}_1, \ldots, \mathbf{u}_k$ are the k orthonormal columns of $\mathbf{U}_k$. Now the

optimization problem becomes

$$\arg\min_{\mathbf{P}\in\mathcal{P}_k} \|\mathbf{PX}-\mathbf{X}\|_F^2 = \arg\max_{\{\mathbf{u}_i\}_{i=1}^k:\text{orthonormal}} \sum_{i=1}^{k} \mathbf{u}_i^T(\mathbf{XX}^T)\mathbf{u}_i.$$

Since $\mathbf{XX}^T$ is SPSD, it has a spectral decomposition

$$\mathbf{XX}^T = \sum_{i=1}^{N} \lambda_i \mathbf{u}_i^* \mathbf{u}_i^{*T},$$

where $\lambda_1 \geq \cdots \geq \lambda_N \geq 0$ are eigenvalues of $\mathbf{XX}^T$ with corresponding orthonormal eigenvectors $\mathbf{u}_1^*, \ldots, \mathbf{u}_N^*$. We now find a unit vector $\mathbf{u}_1$ in $\mathbb{R}^N$ such that $\mathbf{u}_1^T(\mathbf{XX}^T)\mathbf{u}_1$ is maximized. Note that

$$\mathbf{u}_1^T(\mathbf{XX}^T)\mathbf{u}_1 = \sum_{i=1}^{N} \lambda_i \mathbf{u}_1^T \mathbf{u}_i^* \mathbf{u}_i^{*T} \mathbf{u}_1 = \sum_{i=1}^{N} \lambda_i (\mathbf{u}_1^T \mathbf{u}_i^*)^2 \leq \lambda_1 \sum_{i=1}^{N} (\mathbf{u}_1^T \mathbf{u}_i^*)^2 = \lambda_1,$$

since the $N \times N$ identity matrix $\mathbf{I}$ has a spectral decomposition

$\mathbf{I} = \sum_{i=1}^{N} \mathbf{u}_i^* \mathbf{u}_i^{*T}$ and for any unit vector $\mathbf{u}$

$$1 = \mathbf{u}^T\mathbf{u} = \mathbf{u}^T\mathbf{I}\mathbf{u} = \sum_{i=1}^{N} \mathbf{u}^T\mathbf{u}_i^*\mathbf{u}_i^{*T}\mathbf{u} = \sum_{i=1}^{N} (\mathbf{u}^T\mathbf{u}_i^*)^2.$$

By letting $\mathbf{u}_1 = \mathbf{u}_1^*$, the upper bound $\lambda_1 = \mathbf{u}_1^{*T}(\mathbf{X}\mathbf{X}^T)\mathbf{u}_1^*$ is achieved. We next find a unit vector $\mathbf{u}_2$ in $\mathbb{R}^N$ such that $\mathbf{u}_2^T(\mathbf{X}\mathbf{X}^T)\mathbf{u}_2$ is maximized under the constraint that it is orthogonal to $\mathbf{u}_1^*$. Then we have

$$\mathbf{u}_2^T(\mathbf{X}\mathbf{X}^T)\mathbf{u}_2 = \sum_{i=1}^{k} \lambda_i(\mathbf{u}_2^T\mathbf{u}_i^*)^2 = \sum_{i=2}^{k} \lambda_i(\mathbf{u}_2^T\mathbf{u}_i^*)^2 \leq \lambda_2 \sum_{i=2}^{k} (\mathbf{u}_2^T\mathbf{u}_i^*)^2 \leq \lambda_2$$

and the upper bound λ_2 can be achieved by letting $\mathbf{u}_2 = \mathbf{u}_2^*$.

Iteratively, we find a unit vector $\mathbf{u}_j, j \in [3, k]$ in $\mathbb{R}^N$ such that $\mathbf{u}_j^T(\mathbf{X}\mathbf{X}^T)\mathbf{u}_j$ is maximized under the constraint that it is orthogonal to $\mathbf{u}_1^*, \ldots, \mathbf{u}_{j-1}^*$. Then we have

$$\mathbf{u}_j^T(\mathbf{X}\mathbf{X}^T)\mathbf{u}_j = \sum_{i=1}^{k} \lambda_i(\mathbf{u}_j^T\mathbf{u}_i^*)^2 = \sum_{i=j}^{k} \lambda_i(\mathbf{u}_j^T\mathbf{u}_i^*)^2 \leq \lambda_j \sum_{i=j}^{k}(\mathbf{u}_j^T\mathbf{u}_i^*)^2 \leq \lambda_j$$

and the upper bound λ_j can be achieved by letting $\mathbf{u}_j = \mathbf{u}_j^*$. Continuing this process, we find that by letting $\mathbf{u}_1 = \mathbf{u}_1^*$, …, $\mathbf{u}_k = \mathbf{u}_k^*$, the sum $\sum_{i=1}^{k} \mathbf{u}_i^T(\mathbf{X}\mathbf{X}^T)\mathbf{u}_i$ achieves the maximum value $\sum_{i=1}^{k} \lambda_i$. We conclude that $\mathbf{P}^* = \mathbf{U}_k^*\mathbf{U}_k^{*T}$ is a PCA solution. Since $\mathbf{P}^*\mathbf{X} = \mathbf{U}_k^*\mathbf{U}_k^{*T}\mathbf{X}$, $\mathbf{Y} = \mathbf{U}_k^{*T}\mathbf{X}$ is a k-dimensional representation of $\mathbf{X}$, whose reconstruction of $\mathbf{X}$ is $\mathbf{U}_k\mathbf{Y} = \mathbf{P}^*\mathbf{X}$. □

Remarks

- With the PCA solution $\mathbf{P}^* = \mathbf{U}_k^* \mathbf{U}_k^{*T}$, the minimum reconstruction error is

$$\|\mathbf{P}^*\mathbf{X} - \mathbf{X}\|_F^2 = \mathrm{tr}(\mathbf{X}^T\mathbf{X}) - \sum_{i=1}^{k} \lambda_i = \mathrm{tr}(\mathbf{X}\mathbf{X}^T) - \sum_{i=1}^{k} \lambda_i = \sum_{i=k+1}^{N} \lambda_i.$$

- If the rank of the data matrix $\mathbf{X}$ is r (which is also the rank of $\mathbf{X}\mathbf{X}^T$), we have $\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_N$ so that when $k \geq r$, there will be no reconstruction error for any PCA solution $\mathbf{P}^*$.

- PCA can be viewed as projecting the data $\mathbf{X}$ onto the subspace corresponding to the k largest eigenvalues of the covariance matrix $\mathbf{C} = \frac{1}{m}\mathbf{X}\mathbf{X}^T$, i.e., onto the subspace of maximal variance.

The Contents of This Lecture

- Principal component analysis
- Kernel principal component analysis
- Johnson-Lindenstrauss lemma

Transition from Inner Product to Generic PDS Kernel

- $\Phi : \mathscr{I} \to \mathbb{R}^N$: a feature mapping from the input space $\mathscr{I}$ to a feature space $\mathbb{R}^N$.
- $\mathbf{X} = [\Phi(\omega_1), \ldots, \Phi(\omega_m)]$: the $N \times m$ data matrix associated with a random sample $S = (\omega_1, \ldots, \omega_m)$.
- r : the rank of the data matrix $\mathbf{X}$ with $r \leq \min(N, m)$.
- $\mathbf{X} = \mathbf{U}_r^* \mathbf{\Sigma}_r \mathbf{V}_r^{*T}$: a singular value decomposition (SVD) of the data matrix $\mathbf{X}$, where
  - $\mathbf{\Sigma}_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$: the $r \times r$ diagonal matrix of singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ of $\mathbf{X}$.
    - Singular values $\sigma_i, i \in [1, r]$ are the square-roots of nonzero eigenvalues of $\mathbf{X}\mathbf{X}^T$ (and also $\mathbf{X}^T\mathbf{X}$).

- $\mathbf{V}_r^* = [\mathbf{v}_1^*, \ldots, \mathbf{v}_r^*]$: an $m \times r$ matrix consisting of r orthonormal left singular vectors of $\mathbf{X}$ corresponding to singular values $\sigma_1, \ldots, \sigma_r$ respectively.
  - Left singular vectors $\mathbf{v}_i^*, i \in [1, r]$ of $\mathbf{X}$ are orthonormal eigenvectors of $\mathbf{X}^T\mathbf{X}$ corresponding to eigenvalues $\sigma_i^2, i \in [1, r]$ so that

$$\mathbf{X}^T\mathbf{X} = \mathbf{V}_r^*\mathbf{\Sigma}_r^2{\mathbf{V}_r^*}^T.$$

- $\mathbf{U}_r^* = [\mathbf{u}_1^*, \ldots, \mathbf{u}_r^*]$: an $N \times r$ matrix consisting of r orthonormal right singular vectors of $\mathbf{X}$ corresponding to singular values $\sigma_1, \ldots, \sigma_r$ respectively.
  - Right singular vectors $\mathbf{u}_i^*, i \in [1, r]$ of $\mathbf{X}$ are orthonormal eigenvectors of $\mathbf{X}\mathbf{X}^T$ corresponding to eigenvalues $\sigma_i^2, i \in [1, r]$ so that

$$\mathbf{X}\mathbf{X}^T = \mathbf{U}_r^*\mathbf{\Sigma}_r^2{\mathbf{U}_r^*}^T.$$

- $\mathbf{U}_r^* = \mathbf{X}\mathbf{V}_r^*\mathbf{\Sigma}_r^{-1}$: right singular vectors can be obtained from the data matrix, left singular vectors and singular values.
- $\mathbf{K} = [K_{ij}]$: the kernel matrix associated with the sample S with $K_{ij} = \Phi(\omega_i) \cdot \Phi(\omega_j)$.
  - $\mathbf{K} = \mathbf{X}^T\mathbf{X} = \mathbf{V}_r^*\mathbf{\Sigma}_r^2{\mathbf{V}_r^*}^T$.
  - The rank r, singular values, and left singular vectors can be obtained from the kernel matrix $\mathbf{K}$.
- $k \in [1, r]$.
- $\mathbf{U}_k^* = [\mathbf{u}_1^*, \ldots, \mathbf{u}_k^*] = \mathbf{X}\mathbf{V}_r^* \left(\mathbf{\Sigma}_r^{-1}\right)_k$, where $\left(\mathbf{\Sigma}_r^{-1}\right)_k$ is an $r \times k$ matrix consisting of the first k columns of $\mathbf{\Sigma}_r^{-1}$.
- $\mathbf{Y} = {\mathbf{U}_k^*}^T\mathbf{X}$: an optimal k-dimensional representation of the data matrix $\mathbf{X}$ whose reconstruction $\mathbf{U}_k^*\mathbf{Y}$ of $\mathbf{X}$ minimizes the

squared error,

$$\begin{aligned}
\mathbf{Y} &= \left(\boldsymbol{\Sigma}_r^{-1}\right)_k^T \mathbf{V}_r^{*T} \mathbf{X}^T \mathbf{X} = \left(\boldsymbol{\Sigma}_r^{-1}\right)_k^T \mathbf{V}_r^{*T} \mathbf{K} \\
&= \left(\boldsymbol{\Sigma}_r^{-1}\right)_k^T \mathbf{V}_r^{*T} \mathbf{V}_r^* \boldsymbol{\Sigma}_r^2 \mathbf{V}_r^{*T} = \left(\boldsymbol{\Sigma}_r\right)_k^T \mathbf{V}_r^{*T} \\
&= \begin{bmatrix} \sigma_1 \mathbf{v}_1^{*T} \\ \vdots \\ \sigma_k \mathbf{v}_k^{*T} \end{bmatrix}.
\end{aligned}$$

Formulation of Kernel Principal Component Analysis

- $K : \mathscr{I} \times \mathscr{I} \to \mathbb{R}$: a PDS kernel over the input space $\mathscr{I}$.
- $\Phi : \mathscr{I} \to \mathbb{H}$: the feature mapping from $\mathscr{I}$ to the RKHS $\mathbb{H}$ of K.
- $S = (\omega_1, \ldots, \omega_m)$: a random sample of size m drawn i.i.d. from $\mathscr{I}$ according to a distribution D.
- $\mathbf{X} = [\Phi(\omega_1), \ldots, \Phi(\omega_m)]$: the data matrix associated with the random sample S.
- $\mathbf{K} = [K(\omega_i, \omega_j)]$: the $m \times m$ kernel matrix associated with the random sample S.
- $\mathbf{K} = \mathbf{V}^* \mathbf{\Lambda} \mathbf{V}^{*T}$: a spectral decomposition of $\mathbf{K}$ with $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ where $\lambda_1 \geq \cdots \geq \lambda_r > 0 = \lambda_{r+1} = \cdots = \lambda_m$ are eigenvalues of the kernel matrix $\mathbf{K}$.

- k : a fixed integer in $[1, r]$.
- $\mathbf{V}_k^*$: an $m \times k$ matrix consisting of the first k columns of $\mathbf{V}^*$.
- $\sqrt{\mathbf{\Lambda}}_k$: the $k \times k$ main submatrix of $\sqrt{\mathbf{\Lambda}}$.
- $\mathbf{Y} = \sqrt{\mathbf{\Lambda}}_k \mathbf{V}_k^{*T} = \begin{bmatrix} \sqrt{\lambda_1}\mathbf{v}_1^{*T} \\ \vdots \\ \sqrt{\lambda_k}\mathbf{v}_k^{*T} \end{bmatrix}$: an optimal k-dimensional representation of the data matrix $\mathbf{X}$, where $\mathbf{v}_1^*, \ldots, \mathbf{v}_k^*$ are the first k columns of the $m \times m$ orthogonal matrix $\mathbf{V}^*$.

The Contents of This Lecture

- Principal component analysis
- Kernel principal component analysis
- Johnson-Lindenstrauss lemma

Definition: χ^2-Distribution with k Degrees of Freedom

- $X_1, \ldots, X_k$: k independent standard normal random variables with a pdf $f_{X_i}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.
- $X_i^2, i \in [i.k]$: a gamma r.v. with parameters $\lambda = 1/2$ and $\alpha = 1/2$.
  - It has a pdf $f_{X_i^2}(x) = \begin{cases} \frac{e^{-\frac{1}{2}x}}{\sqrt{2\pi x}}, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases}$
  - Its moment generating function is

$$M_{X_i^2}(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha} = \left(\frac{1}{1-2t}\right)^{1/2}.$$

- $Q = X_1^2 + \cdots + X_k^2$: the sum of the squares of k independent standard normal r.v.'s, which is gamma r.v. with parameters $\lambda = 1/2$ and $\alpha = k/2$.

– It has a pdf $f_Q(x) = \begin{cases} \frac{\frac{1}{2}e^{-\frac{1}{2}x}(\frac{1}{2}x)^{\frac{k}{2}-1}}{\Gamma(\frac{k}{2})}, & \text{if } x > 0, \\ 0, & \text{if } x \leq 0. \end{cases}$

– Its moment generating function is

$$M_Q(t) = \left(\frac{1}{1-2t}\right)^{k/2}.$$

- The distribution of Q is called the χ^2-distribution of k degrees of freedom.
  – $E[Q] = \frac{\alpha}{\lambda} = k$.
  – $\text{Var}(Q) = \frac{\alpha}{\lambda^2} = 2k$.

Concentration of χ^2-Distribution with k Degrees of Freedom

Lemma 12.1: Let

- Q : a χ^2-distributed r.v. with k degrees of freedom.

Then, for any $0 < \epsilon < 1$, the following inequality holds:

$$P((1-\epsilon)k \leq Q \leq (1+\epsilon)k) \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}.$$

Proof. By Markov inequality and for $t > 0$, we have

$$P(Q \geq (1+\epsilon)k) = P(e^{tQ} \geq e^{t(1+\epsilon)k}) \leq \frac{E[e^{tQ}]}{e^{t(1+\epsilon)k}} = \frac{(1-2t)^{-k/2}}{e^{t(1+\epsilon)k}}.$$

By taking $t = \frac{\epsilon}{2(1+\epsilon)}$ which is in $(0, \frac{1}{2})$ when ϵ is in $(0, 1)$, we have

$$P(Q \geq (1+\epsilon)k) \leq \left(\frac{1+\epsilon}{e^{\epsilon}}\right)^{k/2} \leq \left(\frac{e^{\epsilon - \frac{\epsilon^2-\epsilon^3}{2}}}{e^{\epsilon}}\right)^{k/2} = e^{-(\epsilon^2-\epsilon^3)k/4}$$

since $1+x \leq e^{x-\frac{x^2-x^3}{2}}$ for all $x \geq 0$. Similarly by Markov inequality and for $t > 0$, we have

$$P(Q \leq (1-\epsilon)k) = P(e^{-tQ} \geq e^{-t(1-\epsilon)k}) \leq \frac{E[e^{-tQ}]}{e^{-t(1-\epsilon)k}} = \frac{(1+2t)^{-k/2}}{e^{-t(1-\epsilon)k}}.$$

By taking $t = \frac{\epsilon}{2(1-\epsilon)}$ which is > 0 when ϵ is in $(0,1)$, we have

$$P(Q \leq (1-\epsilon)k) \leq \left(\frac{1-\epsilon}{e^{-\epsilon}}\right)^{k/2} \leq \left(\frac{e^{-\epsilon-\frac{\epsilon^2-\epsilon^3}{2}}}{e^{-\epsilon}}\right)^{k/2} = e^{-(\epsilon^2-\epsilon^3)k/4}$$

since $1-x \leq e^{-x-\frac{x^2-x^3}{2}}$ for all $x \geq 0$. Thus we have

$$P(Q < (1-\epsilon)k \text{ or } Q > (1+\epsilon)k) \leq 2e^{-(\epsilon^2-\epsilon^3)k/4}$$

so that

$$\begin{aligned} P((1-\epsilon)k \leq Q \leq (1+\epsilon)k) &= 1 - P(Q < (1-\epsilon)k \text{ or } Q > (1+\epsilon)k) \\ &\geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}. \end{aligned}$$

□

Concentration of the Squared Length of a Random Linearly Mapped Vector

Lemma 12.2: Let

- $\mathbf{x} = [x_1, \ldots, x_N]$: a fixed vector in $\mathbb{R}^N$.
- $\mathbf{A}$: a $k \times N$ random matrix whose entries are drawn i.i.d. from $\mathbb{R}$ according to the standard normal distribution $N(0, 1)$.

Then, for any $0 < \epsilon < 1$, the following inequality holds:

$$P\left((1-\epsilon)\|\mathbf{x}\|^2 \leq \|\frac{1}{\sqrt{k}}\mathbf{A}\mathbf{x}\|^2 \leq (1+\epsilon)\|\mathbf{x}\|^2\right) \leq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}.$$

Proof.

- $\mathbf{Y} = [Y_1, \ldots, Y_k]^T \triangleq \mathbf{Ax}$: a random vector of dimension k.
- $Y_i = \sum_{j=1}^{N} A_{ij} x_j, i \in [1, k]$: statistically independent r.v.'s since $A_{ij}, i \in [1, k], j \in [1.N]$, are statistically independent.
- $Y_i = \sum_{j=1}^{N} A_{ij} x_j, i \in [1, k]$: normal r.v.'s since each is a linear combination of independent normal r.v.'s.
- $E[Y_i] = \sum_{j=1}^{N} E[A_{ij}] x_j = 0$.
- $E[Y_i^2] = \sum_{j,k=1}^{N} E[A_{ij} A_{ik}] x_j x_k = \sum_{j}^{N} E[A_{ij}^2] x_j^2 = \|\mathbf{x}\|^2$.
- $Y_1/\|\mathbf{x}\|, \ldots, Y_k/\|\mathbf{x}\|$: i.i.d. r.v.'s with standard normal distribution $N(0, 1)$.
- $Q \triangleq \|\mathbf{A} \frac{\mathbf{x}}{\|\mathbf{x}\|}\|^2 = \|\frac{\mathbf{Y}}{\|\mathbf{x}\|}\|^2 = \sum_{i=1}^{k} \left(\frac{Y_i}{\|\mathbf{x}\|}\right)^2$: the sum of the squares of k independent standard normal r.v.'s, which is χ^2-distributed with k degrees of freedom.

By Lemma 12.1, for any $0 < \epsilon < 1$, the following inequality holds:

$$P((1-\epsilon)k \leq Q \leq (1+\epsilon)k) \geq 1 - 2e^{-(\epsilon^2-\epsilon^3)k/4}.$$

Since $(1-\epsilon)k \leq Q \leq (1+\epsilon)k$ iff $(1-\epsilon)k \leq \|\mathbf{A}\frac{\mathbf{x}}{\|\mathbf{x}\|}\|^2 \leq (1+\epsilon)k$ iff $(1-\epsilon)\|\mathbf{x}\|^2 \leq \|\frac{1}{\sqrt{k}}\mathbf{A}\mathbf{x}\|^2 \leq (1+\epsilon)\|\mathbf{x}\|^2$, the lemma is proved. $\square$

Remark

- When a fixed N-dimensional vector is randomly linearly mapped to the k-dimensional Euclidean space, the squared length of the random vector mapped is sharply concentrated around its mean.

Johnson-Lindenstrauss Lemma

Lemma 12.3: Let

- $\epsilon \in (0, \frac{1}{2}]$: a parameter.
- $m \geq 2$: an integer.
- $k \geq \frac{16 \ln m}{\epsilon^2}$: dimensionality reduction target.

Then for any set V of m points in $\mathbb{R}^N$, there exists a map $f : \mathbb{R}^N \to \mathbb{R}^k$ such that for all $\mathbf{u}, \mathbf{v} \in V$,

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leq \|f(\mathbf{u})-f(\mathbf{v})\|^2 \leq (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2.$$

Proof. We will proceed a probabilistic argument. Consider a mapping $f : \mathbb{R}^N \to \mathbb{R}^k$ which is a linear transformation defined by a $k \times N$ matrix $\frac{1}{\sqrt{k}}\mathbf{A}$:

$$f(\mathbf{x}) \triangleq \frac{1}{\sqrt{k}}\mathbf{A}\mathbf{x},$$

where $\mathbf{A}$ is a random matrix whose entries are drawn i.i.d. from $\mathbb{R}$ according to the standard normal distribution $N(0, 1)$. Let $V^{(2)}$ be the collection of all pairs of distinct vectors in V.

Now we have

$$
\begin{aligned}
& P\left(\bigcap_{\{\mathbf{u},\mathbf{v}\}\in V^{(2)}} \left((1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \le \|f(\mathbf{u}-\mathbf{v})\|^2 \le (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2\right)\right) \\
= \quad & 1 - P\left(\bigcup_{\{\mathbf{u},\mathbf{v}\}\in V^{(2)}} \left((1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \le \|f(\mathbf{u}-\mathbf{v})\|^2 \le (1+\epsilon)\|\mathbf{u}+\mathbf{v}\|^2\right)^c\right) \\
\ge \quad & 1 - \sum_{\{\mathbf{u},\mathbf{v}\}\in V^{(2)}} P\left(\left((1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \le \|f(\mathbf{u}-\mathbf{v})\|^2 \le (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2\right)^c\right) \\
\ge \quad & 1 - \frac{m(m-1)}{2} 2e^{-(\epsilon^2-\epsilon^3)k/4} \text{ by Lemma 12.2.}
\end{aligned}
$$

The existence of an f as stated in the lemma is guaranteed if

$$m(m-1)e^{-(\epsilon^2-\epsilon^3)k/4} < 1 \text{ if and only if } k > \frac{4\ln(m(m-1))}{\epsilon^2(1-\epsilon)}.$$

With $0 < \epsilon \le \frac{1}{2}$, we have $\frac{1}{1-\epsilon} \le 2$ so that the existence of an f is

guaranteed if $k \geq \frac{16 \ln m}{\epsilon^2} > \frac{8 \ln(m(m-1))}{\epsilon^2}$, provided that $m \geq 2$. Note that for any $\mathbf{u}, \mathbf{v} \in V$, we have

$$f(\mathbf{u} - \mathbf{v}) = \frac{1}{\sqrt{k}} \mathbf{A}(\mathbf{u} - \mathbf{v}) = \frac{1}{\sqrt{k}} \mathbf{A}\mathbf{u} - \frac{1}{\sqrt{k}} \mathbf{A}\mathbf{v} = f(\mathbf{u}) - f(\mathbf{v})$$

and the proof is complete. □

Remarks

- The Johnson-Lindenstrauss lemma is a fundamental result in dimensionality reduction which states that any $m \geq 2$ points in a high-dimensional space can be mapped to a much lower dimension, $k = O(\frac{\ln m}{\epsilon^2})$, without distorting pairwise distance between any two points by more than a factor of $(1 \pm \epsilon)$.
- Such a mapping can be found in randomized polynomial time by randomly linearly mapping high-dimensional points to the k-dimensional Euclidean space.