Dimension reduction

October 11, 2020

Abstract

Topics: linear algebra (linear space, basis, projection, eigenvalue and eigenvectors, SVD), PCA, ICA, and LDA.
Readings: Chapter 12 of PRML.

1 Linear algebra

Linear algebra is at the core of dimension reduction and we review the important definitions and properties.

1.1 Linear spaces and basis

$\mathbb{R}^n$ is a vector space containing all n-dimensional vectors $\mathbf{x} = [x_1, \ldots, x_n]^\top$. A linear combination of a set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is defined as

$$\mathbf{v} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \in \mathbb{R}^n, \quad \lambda_i \in \mathbb{R}, i = 1, \ldots, n. \tag{1}$$

A set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly dependent if there is a non-trivial linear combination, such that $\sum_{i=1}^n \lambda_i \mathbf{x}_i = \mathbf{0}$ and $\lambda_i \neq 0$ for some i. If the linear combination $\sum_{i=1}^n \lambda_i \mathbf{x}_i = \mathbf{0}$ implies that all λ_i are zeros, we say $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly independent.

Example 1.1. *$\mathbf{x}_1 = [1, 0]^\top$ and $\mathbf{x}_2 = [0, 1]^\top$ are linearly independent, while $\mathbf{x}_3 = [1, 0]^\top$ and $\mathbf{x}_4 = [2, 0]^\top$ are linearly dependent since the linear combination $-2\mathbf{x}_3 + \mathbf{x}_4 = 0$.*

Definition 1.2. *A basis of $\mathbb{R}^n$ is a set of vectors $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_n\} \subset \mathbb{R}^n$ if B:*

- *are linearly independent;*
- *span $\mathbb{R}^n$: any $\mathbf{x} \in \mathbb{R}^n$ can be expressed as $\mathbf{x} = \sum_{i=1}^n \lambda_i \mathbf{b}_i$ for some real numbers $\lambda_1, \ldots, \lambda_n$.*
- *are minimal: any subset of vectors of B do not span $\mathbb{R}^n$. That is, there is at least one $\mathbf{x} \in \mathbb{R}^n$ that can't be expressed as a linear combination of $\mathbf{b}_1, \ldots, \mathbf{b}_n$.*

The coefficients λ_i, $i = 1, \ldots, n$, when put into a vector $[\lambda_1, \ldots, \lambda_n]^\top$, is called the coordinate vector of $\mathbf{x}$ under the basis B.

Example 1.3. *The standard basis of $\mathbb{R}^2$ is $\mathbf{e}_1 = [0, 1]^\top$ and $\mathbf{e}_2 = [1, 0]^\top$. Another basis of $\mathbb{R}^2$ is $\mathbf{x}_1 = [1, 1]^\top$ and $\mathbf{x}_2 = [1, -1]^\top$.*

The dimension of a vector space V is defined as the number of vectors of a basis of V. The dimension of V will not change when different bases are used. When we talk about dimension reduction in machine learning, we refer to the techniques that embed high-dimension vectors in a lower-dimension vector space.

Definition 1.4. *For two vector spaces $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, a mapping Φ from V to W is linear if for any $\mathbf{x}_1$ and $\mathbf{x}_2$ from V and any $\lambda, \psi \in \mathbb{R}$, we have*

$$\Phi(\lambda \mathbf{x}_1 + \psi \mathbf{x}_2) = \lambda \Phi(\mathbf{x}_1) + \psi \Phi(\mathbf{x}_2). \tag{2}$$

A linear mapping Φ from V to W can be represented as a transform matrix A_Φ that transforms the coordinate vector of an input $\mathbf{x} \in V$ under a basis B of V to the coordinate vector of the output $\mathbf{y} \in W$ under a basis C of W, so that

$$\mathbf{y} = \Phi(\mathbf{x}) = A_\Phi \mathbf{x}. \tag{3}$$

The rank of a matrix $A \in \mathbb{R}^{m \times n}$, denoted by $\text{rank}(A)$, is the maximal number of linearly independent column/row vectors, which ever is smaller. Therefore, $\text{rank}(A) \leq \min\{m, n\}$. Note that the dimension of V may be different from the dimension of W and A does not have to be a square matrix. A is said to be of full rank if $\text{rank}(A) = \min\{m, n\}$.

1.2 Geometry of $\mathbb{R}^n$

Given two non-zero vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$, the angle ω between them is defined by

$$\cos\omega = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \times \|\mathbf{y}\|}. \tag{4}$$

Two non-zero vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* if $\cos\omega = 0$. If in addition $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, we say $\mathbf{x}$ and $\mathbf{y}$ are orthonormal.

Example 1.5. *$\mathbf{x} = [1, 1]^\top$ and $\mathbf{y} = [1, -1]^\top$ are orthogonal ($\langle \mathbf{x}, \mathbf{y} \rangle = 0$) but not orthonormal ($\|\mathbf{x}\| \neq 1$). By scaling these two vectors, $\mathbf{x} = \frac{1}{\sqrt{2}}[1, 1]^\top$ and $\mathbf{y} = \frac{1}{\sqrt{2}}[1, -1]^\top$ are orthonormal.*

A square matrix $A \in \mathbb{R}^{n \times n}$ is called an orthogonal matrix if the columns vectors and row vectors are orthonormal[1] so that

$$I = A^\top A = A A^\top. \tag{5}$$

In other words, the inverse matrix of A, denoted by A^{-1}, is $A^\top$.

An orthogonal matrix represents a linear mapping that is a rotaion: $A\mathbf{x}$ rotates $\mathbf{x}$. Therefore, as a mapping, the orthogonal matrix preserves lengths of vectors and angles between any two vectors:

$$\|A\mathbf{x}\|^2 = (A\mathbf{x})^\top (A\mathbf{x}) = \mathbf{x}^\top A^\top A \mathbf{x} = \mathbf{x}^\top I \mathbf{x} = \|\mathbf{x}\|^2, \tag{6}$$

$$\cos\omega = \frac{\langle A\mathbf{x}, A\mathbf{y} \rangle}{\|A\mathbf{x}\| \times \|A\mathbf{y}\|} = \frac{\mathbf{x}^\top A^\top A \mathbf{y}}{\sqrt{(\mathbf{x}^\top A^\top A \mathbf{x})(\mathbf{y}^\top A^\top A \mathbf{y})}} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{x}\| \times \|\mathbf{y}\|}.$$

A basis $B = [\mathbf{b}_1, \ldots, \mathbf{b}_n]$ of the vector space $\mathbb{R}^n$ is an OrthoNormal Basis (ONB) if the matrix B is orthogonal:

$$BB^\top = B^\top B = I. \tag{7}$$

For example, $\mathbf{b}_1 = \frac{1}{\sqrt{2}}[1, 1]^\top$ and $\mathbf{b}_2 = \frac{1}{\sqrt{2}}[1, -1]^\top$ is a orthonormal basis of $\mathbb{R}^2$.

The following definition will be critical to PCA.

Definition 1.6. *Given an M-dimension subspace U of the vector space $V = \mathbb{R}^n$, the orthogonal complement of U, denoted by $U^\perp$, is a $D = n - M$ dimension subspace of V, such that any vector in $U^\perp$ is orthogonal to any vector in U.*

The only vector in the intersection of U and $U^\perp$ is the zero vector $\mathbf{0}$. As a result, any vector $\mathbf{x} \in V$ can be uniquely represented as

$$\mathbf{x} = \sum_{j=1}^{M} \lambda_j \mathbf{b}_j + \sum_{j=1}^{D} \lambda_j \mathbf{b}_j^\perp, \tag{8}$$

where $(\mathbf{b}_1, \ldots, \mathbf{b}_M)$ is the M basis vectors for U and $(\mathbf{b}_1^\perp, \ldots, \mathbf{b}_D^\perp)$ is the D basis vectors for $U^\perp$.

[1] A should have been called "orthonormal matrix" but we follow the convention in linear algebra

1.3 Projection

We have seen projection in the SVM lectures, where we project an example $\mathbf{x}$ to the normal vector of a hyperplane that separates the positive and negative training data. PCA can be regarded as projecting data in a high dimension space to a lower dimension space. We discuss more general projection.

Definition 1.7. *Let V be a vector space and $U \subset V$ a subspace of V. A linear mapping $\pi : V \to U$ is called a projection if $\pi^2 \triangleq \pi \circ \pi = \pi$.*

That is, if we project a vector $\mathbf{x} \in V$ to U using π, a further projection using π will not change the first projection. Since we work on $V = \mathbb{R}^n$, U is a subspace of $\mathbb{R}^n$ and we can use the projection matrix P_π associated with π to represent the linear mapping π. That is, $\pi(\mathbf{x}) = P_\pi \mathbf{x}$. By definition, $P_\pi^2 = P_\pi P_\pi = P_\pi$.

In the SVM lectures, the projection of $\mathbf{x} \in \mathbb{R}^n$ onto a non-zero single vector $\mathbf{w}$ (which is the basis of a 1-dimension subspace of $\mathbb{R}^n$) is

$$\frac{\mathbf{x}^\top \mathbf{w}}{\|\mathbf{w}\|^2}\mathbf{w} = \frac{\mathbf{w}(\mathbf{x}^\top \mathbf{w})}{\|\mathbf{w}\|^2} = \frac{\mathbf{w}(\mathbf{w}^\top \mathbf{x})}{\|\mathbf{w}\|^2} = \frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}\mathbf{x} \tag{9}$$

Therefore, the projection matrix is $P_\mathbf{w} = \frac{\mathbf{w}\mathbf{w}^\top}{\|\mathbf{w}\|^2}$. We can verify that $P_\mathbf{w}^2 \mathbf{x} = P_\mathbf{w} \mathbf{x}$.

Example 1.8. *Project $\mathbf{x} = [1, 1, 1]^\top \in \mathbb{R}^3$ onto the 1 dimension vector space U spanned by $\mathbf{b} = [1, 2, 2]^\top$. The projection matrix is $\mathbf{b}\mathbf{b}^\top / \|\mathbf{b}\|^2$ and the projection is $\mathbf{b}\mathbf{b}^\top \mathbf{x} / \|\mathbf{b}\|^2$.*

In PCA, we will pursue a more general projection that maps $\mathbf{x} \in V$ to a subspace U that can have more than 1 dimension. Assume that U is of dimension $m \geq 1$ with a basis $B = [\mathbf{b}_1, \ldots, \mathbf{b}_m] \in \mathbb{R}^{n \times m}$. Any projection of $\mathbf{x} \in \mathbb{R}^n$ onto U, denoted by $\pi_U(\mathbf{x})$, can be represented by

$$\pi_U(\mathbf{x}) = \sum_{j=1}^{m} \lambda_j \mathbf{b}_j = B\boldsymbol{\lambda}, \quad \boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^\top \in \mathbb{R}^m. \tag{10}$$

The vector pointing from the projection $\pi_U(\mathbf{x})$ to $\mathbf{x}$ must be orthogonal (perpendicular) to any vector in U. This is equivalent to the set of m linear equations of $\boldsymbol{\lambda}$:

$$\langle \mathbf{b}_i, \mathbf{x} - \pi_U(\mathbf{x}) \rangle = \mathbf{b}_i^\top (\mathbf{x} - B\boldsymbol{\lambda}) = 0, \quad i = 1, \ldots, m. \tag{11}$$

Written in matrix form, we have

$$B^\top (\mathbf{x} - B\boldsymbol{\lambda}) = \mathbf{0} \tag{12}$$

Solving for $\boldsymbol{\lambda}$, we obtain

$$\boldsymbol{\lambda} = (B^\top B)^{-1} B^\top \mathbf{x}. \tag{13}$$

Remark: the inverse $(B^\top B)^{-1}$ exists since B is assumed to contain the m basis vectors as columns and the rank of B is m (namely, B is full rank). Eq. (13) is called the "normal equation". The closed-form solution to the linear regression problem is a special case of normal equations, with $B = X \in \mathbb{R}^{m \times n}$ being the design matrix, $\mathbf{x} = \mathbf{y} \in \mathbb{R}^m$ being the vector of m target values, and $\boldsymbol{\lambda}$ being the fitted weight vector $\mathbf{w} \in \mathbb{R}^n$.

Plugging $\boldsymbol{\lambda}$ into Eq. (10), we obtain

$$\pi_U(\mathbf{x}) = B(B^\top B)^{-1} B^\top \mathbf{x} = P_U \mathbf{x}, \tag{14}$$

so that the projection matrix is $P_U = B(B^\top B)^{-1} B^\top$. It can be verified that $P_U^2 = P_U$.

Example 1.9. *Try to use Eq. (14) to compute the projection of $\mathbf{x} = [6, 0, 0]^\top \in \mathbb{R}^3$ onto the 2 dimension vector space U spanned by $\mathbf{b}_1 = [1, 1, 1]^\top$ and $\mathbf{b}_2 = [0, 1, 2]^\top$.*

Remark: 1) the projection $\pi_U(\mathbf{x})$ is still a vector in $\mathbb{R}^n$, but it can be represented by only m coefficients under the basis $B = [\mathbf{b}_1, \ldots, \mathbf{b}_m]$ of U of dimension m. That is, the coefficient vector of $\pi_U(\mathbf{x})$ under the basis B is only in dimension m. 2) If B is a orthogonal matrix, that is, $B^\top B = I$, then the projection can be simplified to $\pi_U(\mathbf{x}) = BB^\top \mathbf{x}$.

1.4 Eigenvalues and eigenvectors

Eigenvalues and eigenvectors reveal inherent properties of square matrices, and the singular value decomposition (SVD) deals with the more general matrices that may not be square matrices. These techniques are examples of *matrix factorization*

$$A = BC, \tag{15}$$

similar to factorization of an integer (e.g., 10) into the product of multiple integers (2×5).

Definition 1.10. *Given a square matrix $A \in \mathbb{R}^{n\times n}$, then $\lambda \in \mathbb{R}$ is an eigenvalue of A and $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ is the corresponding eigenvector of A if*

$$A\mathbf{x} = \lambda\mathbf{x}. \tag{16}$$

Eq. (16) is called *eigenvalue equation*.

Remark: 1) there can be multiple eigenvalues and eigenvectors for a square matrix A; 2) an eigenvalue can be repeated more than once with different eigenvectors, which forms the basis vectors of a subspace (eigenspace) of $\mathbb{R}^n$; 3) the pair $(\lambda, \mathbf{x})$ can be scaled to $(c\lambda, c\mathbf{x})$ by a constant c as another pair of eigenvalue and eigenvector, since $A(c\mathbf{x}) = cA\mathbf{x} = c\lambda\mathbf{x}$; 4) if A has n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$, then the corresponding eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are linearly independent (forming a basis for $\mathbb{R}^n$).

Let $P = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ be the matrix with eigenvectors as columns, and D is an $n \times n$ diagonal matrix with diagonal elements being the corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$.

$$AP = [A\mathbf{x}_1, \ldots, A\mathbf{x}_n] = PD = [\lambda_1\mathbf{x}_1, \ldots, \lambda_n\mathbf{x}_n], \tag{17}$$

which is just another way to say that $A\mathbf{x}_i = \lambda_i\mathbf{x}_i$ for $i = 1, \ldots, n$.

For a general square matrix A, it may not have n linearly independent eigenvectors and the inverse of P may not exist. When P^{-1} does exist, $A^{-1}A = I$ and $AP = PD$ can be re-written as

$$A = PDP^{-1} \tag{18}$$

and we call this decomposition of A the "eigen-decomposition" and A is similar to a diagonal matrix with eigenvalues (i.e., A is "diagnolizable").

For symmetric matrices, the eigen-decomposition always exist according to the following theorem.

Theorem 1.11. *Spectral theorem: If $A \in \mathbb{R}^{n\times n}$ is symmetric, there exists an orthonormal basis of the corresponding vector space V consisting of eigenvectors of A, and each eigenvalue is real.*

As a result, any symmetric matrix A can be factorized as

$$A = PDP^\top, \tag{19}$$

Since an orthogonal matrix is a rotation, the eigen-decomposition of a symmetric matrix A represent A as a sequence of three linear mappings: first $P^\top$ rotates $\mathbf{x}$, then D scales $P^\top\mathbf{x}$, and lastly P rotates $DP^\top\mathbf{x}$ to $PDP^\top\mathbf{x} = A\mathbf{x}$.

1.5 Singular Value Decomposition (SVD)

Eigen-decomposition works only for square matrices and singular value decomposition (SVD) generalizes the decomposition to non-square matrices. Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank $r \leq \min\{m, n\}$. The SVD of A is a decomposition of the form

$$A = U\Sigma V^\top, \tag{20}$$

where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_m] \in \mathbb{R}^{m\times m}$ is an orthogonal matrix called the "left-singluar-matrix", $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{n\times n}$ is an orthogonal matrix called the "right-singluar-matrix", and the $m \times n$ matrix Σ with singular values $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{ij} = 0$ for $i \neq j$.

Example 1.12. *The following example is taken from Example 4.12 of the book "Mathematics for Machine Learning". Let the matrix*

$$A = \begin{bmatrix} 1 & -0.8 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \tag{21}$$

The SVD of A is

$$A = U\Sigma V^\top = \begin{bmatrix} -0.79 & 0 & -0.62 \\ 0.38 & -0.78 & -0.49 \\ -0.48 & -0.62 & 0.62 \end{bmatrix} \begin{bmatrix} 1.62 & 0 \\ 0 & 1.0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.78 & 0.62 \\ -0.62 & -0.78 \end{bmatrix} \tag{22}$$

Similar to eigendecomposition, SVD decomposes a linear mapping A into a sequence of three linear mappings: first the orthogonal matrix $V^\top \in \mathbb{R}^n$ rotates $\mathbf{x} \in \mathbb{R}^n$, then Σ scales $V^\top \mathbf{x}$ along each of the n axes. If $m \geq n$, Σ is a tall matrix that places the scaled vector in a n dimension subspace of the m dimension space, and if $m \leq n$, Σ is a wide matrix that truncates the last $n-m$ rows of $V^\top \mathbf{x}$. Lastly, $U \in \mathbb{R}^{m \times m}$ rotates $\Sigma V^\top \mathbf{x}$ in $\mathbb{R}^m$.

Properties of SVD:

- If A is a symmetric matrix, the eigen-decomposition and SVD of A are identical.
- $U(\Sigma\Sigma^\top)U^\top = AA^\top$ is the eigen-decomposition of $AA^\top$.
- $V(\Sigma\Sigma^\top)V^\top = A^\top A$ is the eigen-decomposition of $A^\top A$.

2 Principal Component Analysis

Assume that we are given data $\{\mathbf{x}^{(i)}\}_{i=1}^m$, with $\mathbf{x}^{(i)} \in \mathbb{R}^n$. Let the design matrix $X = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}] \in \mathbb{R}^{n \times m}$. We assume that the data are centered so that $\frac{1}{m}\sum_{i=1}^m \mathbf{x}^{(i)} = 0$. Typically n is a rather large number. For example, if each $\mathbf{x}^{(i)}$ is a vector representation of pixels of an image, then n can easily to up to hundreds of thousands; if each $\mathbf{x}^{(i)}$ is a vector representing a document, n typically are in millions. Among these all dimensions, one may be related to another dimension, and some dimensions can not capture data variance. Principal Component Analysis (PCA) is a dimension reduction technique that maps X to $\tilde{X} \in \mathbb{R}^{d \times m}$, so that each $\tilde{\mathbf{x}}^{(i)}$ is in a lower dimension vector space $\mathbb{R}^d$, $d < n$. $\tilde{\mathbf{x}}^{(i)}$ are linear combinations of $\mathbf{x}^{(i)}$, preserving the data variance while minimizing the loss in information due to the reduction of features. There are two formulations of PCA.

2.1 Maximum variance formulation

Let's consider the special case when $d = 1$. Let the linear mapping be $\mathbf{u}_1$ that maps $\mathbf{x}^{(i)}$ to $\mathbf{u}_1\mathbf{x}^{(i)} \in \mathbb{R}$. Then the variance of the mapped data is

$$J(\mathbf{u}) = \frac{1}{m}\sum_{i=1}^m (\mathbf{u}_1^\top \mathbf{x}^{(i)} - \mathbb{E}[\mathbf{u}_1^\top \mathbf{x}^{(i)}])^2 = \frac{1}{m}\sum_{i=1}^m (\mathbf{u}_1^\top \mathbf{x}^{(i)})^2. \tag{23}$$

since the data are centered and $\mathbb{E}[\mathbf{u}_1^\top \mathbf{x}^{(i)}] = 0$. Using symmetry of inner product,

$$J(\mathbf{u}) = \frac{1}{m}\sum_{i=1}^m (\mathbf{u}_1^\top \mathbf{x}^{(i)})(\mathbf{u}_1^\top \mathbf{x}^{(i)}) = \frac{1}{m}\sum_{i=1}^m (\mathbf{u}_1^\top \mathbf{x}^{(i)})({\mathbf{x}^{(i)}}^\top \mathbf{u}_1) = \mathbf{u}_1^\top \frac{1}{m}\sum_{i=1}^m (\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^\top)\mathbf{u}_1 \tag{24}$$

Let $S = \frac{1}{m}\sum_{i=1}^m (\mathbf{x}^{(i)}{\mathbf{x}^{(i)}}^\top)$ be the covariance matrix. The maximization of the function $J(\mathbf{u}_1)$ with respect to $\mathbf{u}_1$ is subject to the constraint $\|\mathbf{u}_1\|^2 = 1$ (since scaling $\mathbf{u}_1$ will also scale J quadratically). Let the Lagrangian function be

$$\mathbf{u}_1^\top S\mathbf{u}_1 + \lambda(1 - \|\mathbf{u}_1\|). \tag{25}$$

Let the partial derivative of the Lagrangian with respect to $\mathbf{u}_1$ be zero, we obtain

$$S\mathbf{u}_1 - \lambda\mathbf{u}_1 \tag{26}$$

We recognize that $\mathbf{u}_1$ is an eigenvector of the symmetric matrix S and λ is the corresponding eigenvalue. Then the objective function J becomes

$$J(\mathbf{u}) = \mathbf{u}_1^\top S\mathbf{u}_1 = \lambda\|\mathbf{u}_1\|^2 = \lambda. \tag{27}$$

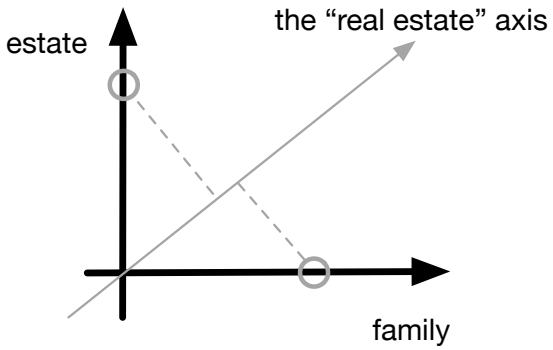

Figure 1: If two documents contains keywords that do not overlap, the similarity between the two documents will be zero. PCA can find a direction to which documents can be projected and similarity can be measured by their semantics about real estate.

By selecting the eigenvector that corresponding to the largest eigenvalue, denoted by λ_1, we find the first direction to project the data.

One projection may not sufficiently preserve the variance in the data and adding more projection direction is desirable. Pick the next $d-1$ eigenvectors $\mathbf{u}_2, \ldots, \mathbf{u}_d$ that corresponding to the next $d-1$ largest eigenvalues $\lambda_1, \ldots, \lambda_d$, we can find the projection coefficients of $\mathbf{x}^{(i)}$ onto the top d eigenvectors

$$\mathbf{z}^{(i)} = U\mathbf{x}^{(i)} = [\mathbf{u}_1^\top \mathbf{x}^{(i)}, \ldots, \mathbf{u}_d^\top \mathbf{x}^{(i)}] \in \mathbb{R}^d. \tag{28}$$

where $U = [\mathbf{u}_2, \ldots, \mathbf{u}_d] \in \mathbb{R}^{n \times d}$ is a basis of the subspace $\mathbb{R}^d$ of $\mathbb{R}^n$. Therefore, we encode $\mathbf{x}^{(i)}$ using d coefficients rather than n ones. The code $\mathbf{z}^{(i)}$ helps reconstruct the projection $\tilde{\mathbf{x}}^{(i)}$ as an approximation of $\mathbf{x}^{(i)}$

$$\tilde{\mathbf{x}}^{(i)} = UU^\top \mathbf{x}^{(i)}. \tag{29}$$

2.2 Minimum error formulation

Let $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ be a set of orthonormal basis of the vector space $\mathbb{R}^n$.

$$\mathbf{x}^{(i)} = UU^\top \mathbf{x}^{(i)} = \sum_{j=1}^{n} \mathbf{u}_j (\mathbf{u}_j^\top \mathbf{x}^{(i)}). \tag{30}$$

Suppose we only want to keep the first $n-1$ component in the above summation, then the error in approximating all $\mathbf{x}^{(i)}$ is

$$J(\mathbf{u}_n) = \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{x}^{(i)} - \sum_{j=1}^{n-1} \mathbf{u}_j (\mathbf{u}_j^\top \mathbf{x}^{(i)})\|^2 = \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{u}_n (\mathbf{u}_n^\top \mathbf{x}^{(i)})\|^2 = \frac{1}{m} \sum_{i=1}^{m} (\mathbf{u}_n^\top \mathbf{x}^{(i)})^2 = \mathbf{u}_n^\top S \mathbf{u}_n$$

We can see that by selecting the eigenvector for $\mathbf{u}_n$ that corresponding to the smallest eigenvalue λ_n of S will minimize the above objective. In other words, we can pick the top $n-1$ eigenvectors to project the data and minimize the approximation error. In general, less than $n-1$, say d, eigenvectors can be selected and the approximation error is $\sum_{j=d+1}^{n} \lambda_j$.

2.3 Applying PCA to text data

In search engines, a user issues a query, such as “real estate”, and the search engine should return a set of documents related to real estate, containing keywords such as “three bedrooms”, “single family”, etc. The returned documents may not contain the queried keyword and the similarity between the query and a document will be zero (“real estate”$\neq$“single familry”). The search engine needs to infer the semantic similarity between the query and the documents to find good semantic matches. In other words, the matching happens on the semantic level (“real estate”$\sim$“single familry”) rather than on the vocabulary level.

PCA can project text documents to semantic dimensions so that text similarity cna be measured on the semantic dimensions and more accurately model the contents. The method is also called Latent Semantic Indexing (LSI). An example is shown in Figure 1.

3 Supervised dimension reduction

PCA processes unlabeled data and maximize the variance of the projected data. When labels of the data are given, a projection that preserves the information in the data labels is desirable.

Let the training data be $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$. Each instance $\mathbf{x}^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \{0, 1\}$. Let $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ be grouped by their labels so that $\mathcal{C}_0 = \{\mathbf{x}^{(i)} : y^{(i)} = 0\}$ and $\mathcal{C}_1 = \{\mathbf{x}^{(i)} : y^{(i)} = 1\}$ are groups of negative and positive training examples. More than two groups can be handled (see PRML 4.1.6). Let the mean vectors of the two groups be

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{\mathbf{x} \in \mathcal{C}_k} \mathbf{x}, \quad k = 0, 1. \tag{31}$$

We aim to find a direction $\mathbf{w}$, so that the projection of the positive and negative groups onto $\mathbf{w}$ will be separated with the maximal distance. One possible choice of $\mathbf{w}$ is $\mathbf{m}_0 - \mathbf{m}_1$, that maximizing $\mathbf{w}^\top(\mathbf{m}_0 - \mathbf{m}_1)$. But this is problematic: if the data within the two groups are widely spread out, then the mean vectors can't represent $\mathcal{C}_0$ and $\mathcal{C}_1$ well and there can be a significant overlap between the two projected groups. See Figure 4.6 of PRML.

3.1 Fisher Linear Discriminant Analysis

The above problem can be addressed by additionally minimizing the "scatter" of the projected data within each group, so that the overlap between groups can be reduced. Formally, the mean of the projected groups are

$$m_k = \mathbf{w}^\top \mathbf{m}_k, \quad k = 0, 1, \tag{32}$$

where m_k is a scalar. Similarly, the projection of $\mathbf{x}$ onto $\mathbf{w}$ is $y = \mathbf{w}^\top \mathbf{x}$. The distance between the two groups after the projection is then

$$(m_0 - m_1)^2 = (\mathbf{w}^\top(\mathbf{m}_0 - \mathbf{m}_1))(\mathbf{w}^\top(\mathbf{m}_0 - \mathbf{m}_1)) = \mathbf{w}^\top(\mathbf{m}_0 - \mathbf{m}_1)(\mathbf{m}_0 - \mathbf{m}_1)^\top \mathbf{w} = \mathbf{w}^\top S_B \mathbf{w}, \tag{33}$$

where

$$S_B = (\mathbf{m}_0 - \mathbf{m}_1)(\mathbf{m}_0 - \mathbf{m}_1)^\top \in \mathbb{R}^{n \times n} \tag{34}$$

is the "*between-class scatter matrix*", representing the variance in the two means m_0 and m_1.

The variance of the projected data within class k is

$$\sum_{\mathbf{x} \in \mathcal{C}_k} (\mathbf{w}^\top \mathbf{x})^2 = \sum_{\mathbf{x} \in \mathcal{C}_k} (\mathbf{w}^\top \mathbf{x})(\mathbf{x}^\top \mathbf{w}) = \mathbf{w}^\top \left(\sum_{\mathbf{x} \in \mathcal{C}_k} \mathbf{x}\mathbf{x}^\top \right) \mathbf{w} = \mathbf{w}^\top S_k \mathbf{w}, \tag{35}$$

where

$$S_k = \sum_{\mathbf{x} \in \mathcal{C}_k} \mathbf{x}\mathbf{x}^\top \tag{36}$$

represents how much variance there are in the projected data from group k. Let the "*within-class scatter matrix*" be $S_W = S_1 + S_2$ and $\mathbf{w}^\top S_W \mathbf{x}$ represents the overall variances of the projected data from both groups. For simplicity, we assume both S_B and S_W are positive definite so that all their eigenvalues are positive.

To maximize the between group variance and minimize the within group varaince, we minimize the following objective function

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}} \tag{37}$$

Note that both the numerator and denominator are scalars. The ratio is called the "Rayleigh ratio".

There are two ways to solve for the optimal $\mathbf{w}$.

- Note that the scale of $\mathbf{w}$ does not matter and only the direction matters (scaling $\mathbf{w}$ to $\alpha\mathbf{w}$ will not change the objective function value), and we can restrict $\mathbf{w}^\top S_W \mathbf{w} = 1$. The above problem becomes a constrained optimization problem

$$\begin{aligned} &\min_{\alpha_2} && -\mathbf{w}^\top S_B \mathbf{w} \\ &\text{s.t.} && \mathbf{w}^\top S_W \mathbf{w} = 1 \end{aligned} \tag{38}$$

Construct the Lagrangian function

$$L(\mathbf{w}, \lambda) = -\mathbf{w}^\top S_B \mathbf{w} + \lambda(\mathbf{w}^\top S_W \mathbf{w} - 1). \tag{39}$$

Take the partial derivative of L with respect to $\mathbf{w}$, we obtain

$$S_B \mathbf{w} = \lambda S_W \mathbf{w}. \tag{40}$$

Since S_B is symmetric, according to the spectral theorem, there is a eigen-decomposition $S_B = U\Lambda U^\top$, where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ are the eigenvectors and Λ is a diagonal matrix with eigenvalues on the diagonal. $S_B^{1/2} = U\Lambda^{1/2}U^\top$ with $U^{1/2}$ being a diagonal matrix with the square root of the eigenvalues on the diagonal.

$$S_W^{-1} S_B^{1/2} S_B^{1/2} \mathbf{w} = \lambda S_B^{-1/2} S_B^{1/2} \mathbf{w}. \tag{41}$$

Let $S_B^{1/2}\mathbf{w} = \mathbf{v}$, then

$$S_B^{1/2} S_W^{-1} S_B^{1/2} \mathbf{v} = \lambda \mathbf{v}. \tag{42}$$

This is the eigenvalue equation for the symmetric matrix $S_B^{1/2} S_W^{-1} S_B^{1/2}$. The remaining question is which eigenvector $\mathbf{v}$ of $S_B^{1/2} S_W^{-1} S_B^{1/2}$ to pick.

Plugging $\mathbf{w} = S_B^{-1/2}\mathbf{v}$ into the Rayleigh ratio in Eq. (37),

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top S_B \mathbf{w}}{\mathbf{w}^\top S_W \mathbf{w}} = \frac{\mathbf{v}^\top \mathbf{v}}{\mathbf{v}^\top S_B^{-1/2} S_W S_B^{-1/2} \mathbf{v}} = 1/(1/\lambda) = \lambda. \tag{43}$$

To maximize the ratio, select the eigenvector that has the largest eigenvalue.

- An easier way is to take partial derivative of $J(\mathbf{w})$ and set the derivative to $\mathbf{0}$. Then

$$(\mathbf{w}^\top S_w \mathbf{w}) S_B \mathbf{w} = (\mathbf{w}^\top S_B \mathbf{w}) S_W \mathbf{w}. \tag{44}$$

Since $S_B = (\mathbf{m}_0 - \mathbf{m}_1)(\mathbf{m}_0 - \mathbf{m}_1)^\top$, $S_B\mathbf{w} = (\mathbf{m}_0 - \mathbf{m}_1)(\mathbf{m}_0 - \mathbf{m}_1)^\top \mathbf{w} = \lambda(\mathbf{m}_0 - \mathbf{m}_1)$. Therefore, we find that

$$\mathbf{w} = \lambda' S_W^{-1}(\mathbf{m}_0 - \mathbf{m}_1). \tag{45}$$

for some scalar λ'.