机器学习引论

彭玺

pengxi@scu.edu.cn

www.pengxi.me

提纲

一．Review

二．Dimension Reduction

三．Principle Component Analysis

提纲

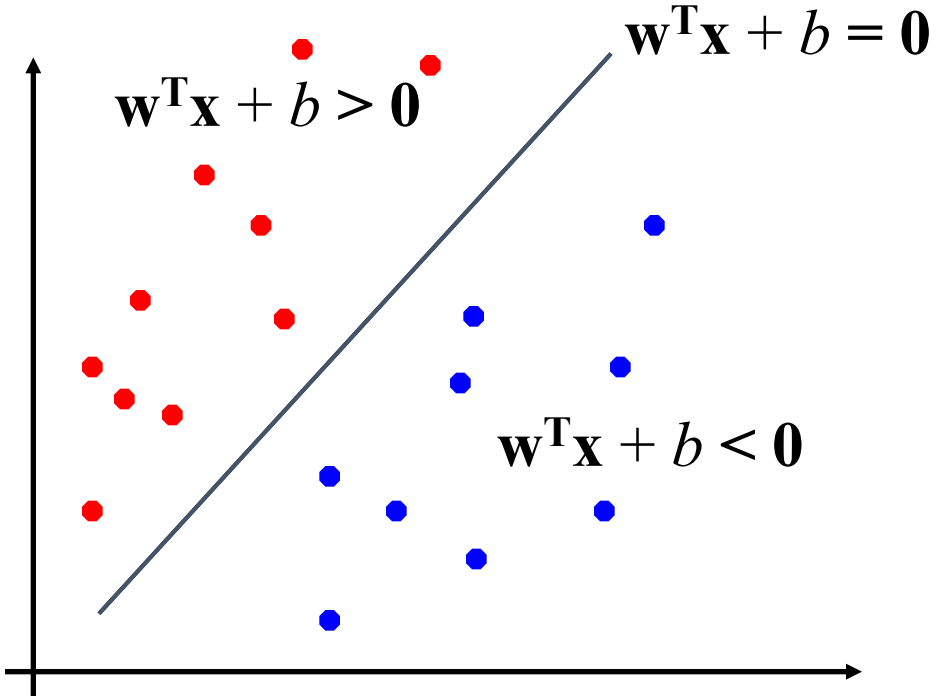
一．Review

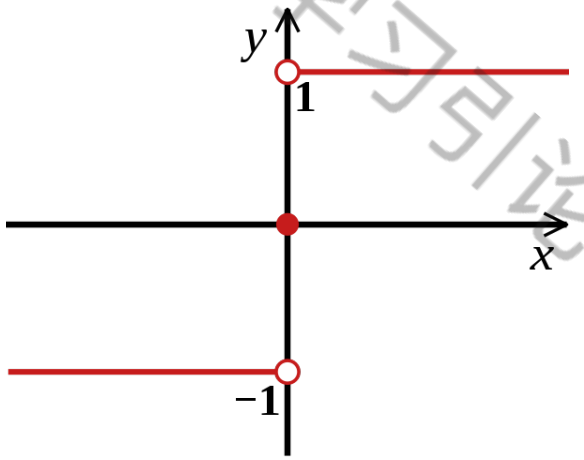
二．Dimension Reduction

三．Principal Component Analysis

Review

- Binary classification can be viewed as the task of separating classes in feature space:



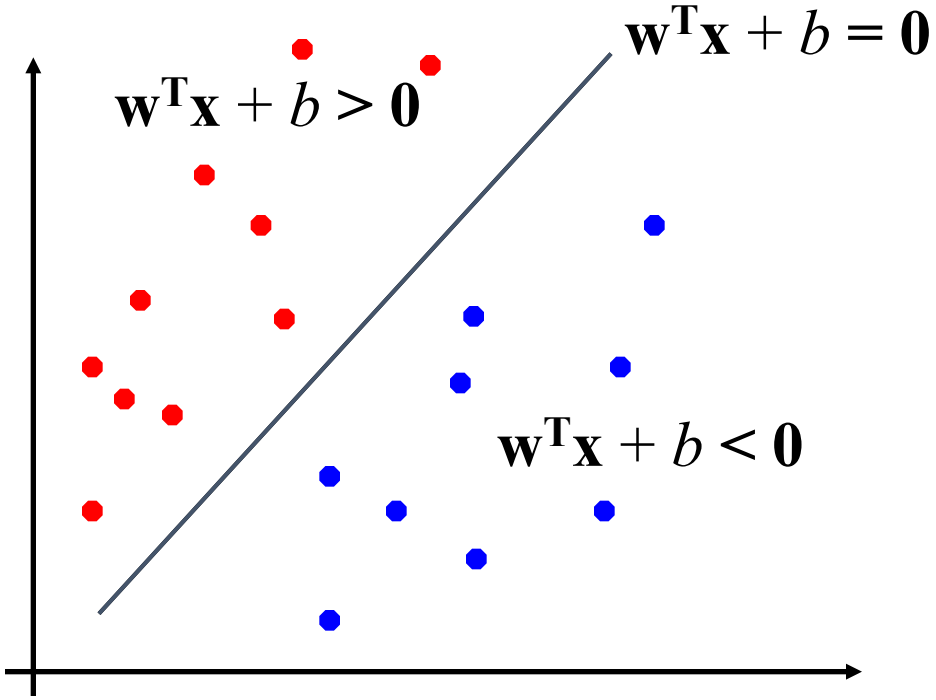


$$\operatorname{sgn}(x) := \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$$

Activate function

Review

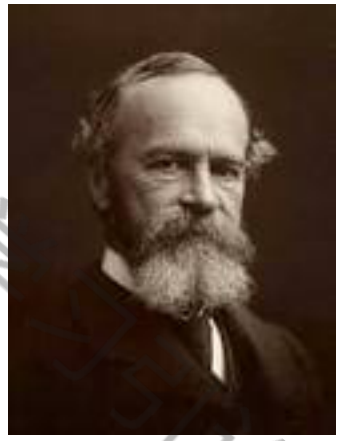
- Binary classification can be viewed as the task of separating classes in feature space:

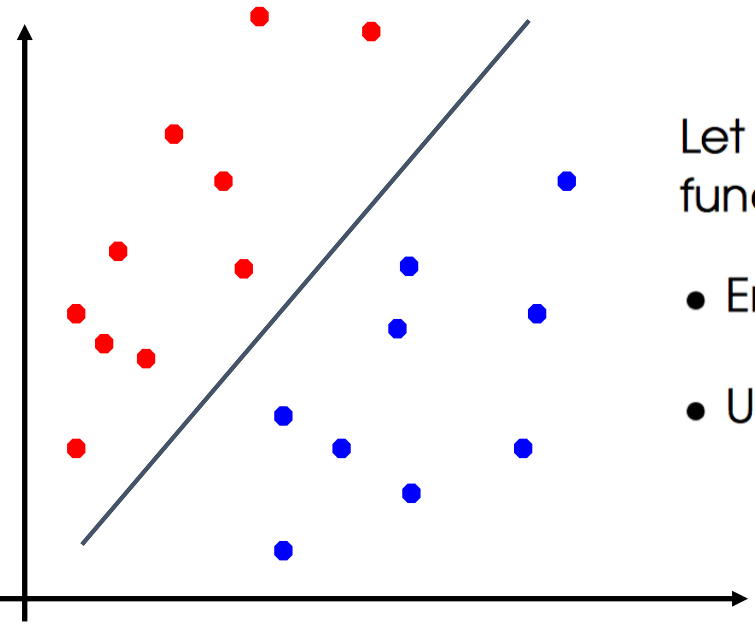


$$y = \mathrm{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x} + b)$$

Review

- 1890：美国心理学家和哲学家William James在其著作中指出——当两个事件同时发生时，涉及到的大脑过程间的连接将会增强，这是无监督的Hebb学习规则的灵感来源。此外，James还提出了加权（weighted）、可变（modifiable）、及并行连接（parallel connections）等神经网络至今采用的基本概念。





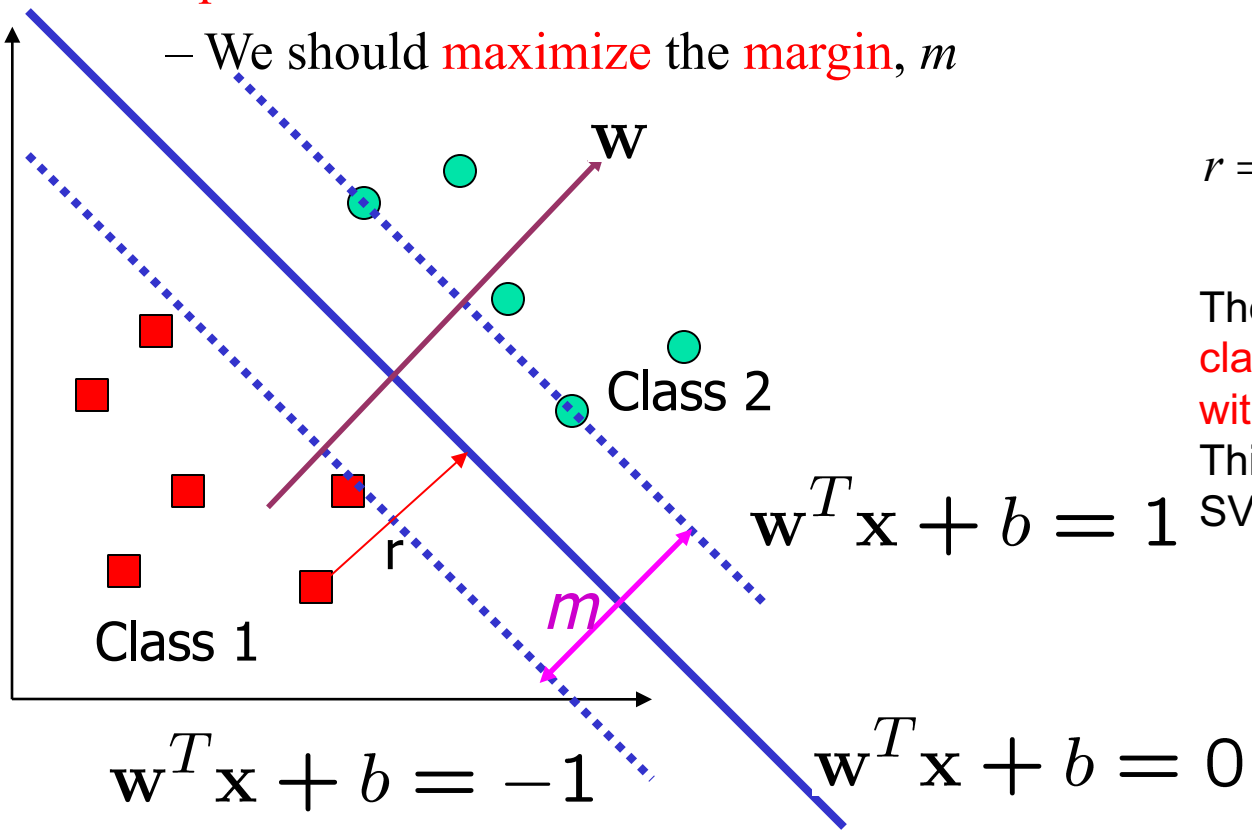
Let y be the correct output, and $f(x)$ the output function of the network.

- Error: $\mathrm{E} = y - f(x)$
- Update weights: $W_j \leftarrow W_j + \alpha x_j E$

Review

The decision boundary should be as far away from the data of both classes as possible

– We should maximize the margin, m



$$r=\frac{\mathbf{w}^T\mathbf{x}_i+b}{\|\mathbf{w}\|}$$

The maximum margin linear classifier is the linear classifier with the maximum margin. This is the simplest kind of SVM (Called an Linear SVM)

Review

- Let training set $\{(\mathbf{x}_i, y_i)\}_{i=1..n}$, $\mathbf{x}_i \in \mathrm{R}^d$, $y_i \in \{-1, 1\}$ be separated by a hyperplane with margin m. Then for each training example $(\mathbf{x}_i, y_i)$:

$$\begin{aligned} \mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b &\leq -m/2 \quad \text{if } y_i = -1 \\ \mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b &\geq m/2 \quad \text{if } y_i = 1 \end{aligned} \quad \Leftrightarrow \quad y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b) \geq m/2$$

- For every support vector $\mathbf{x}_s$, the above inequality is an equality. After rescaling $\mathbf{w}$ and b by $m/2$ in the equality, we obtain that distance between each $\mathbf{x}_s$ and the hyperplane is $r = \frac{y_s(\mathbf{w}^T\mathbf{x}_s + b)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$

- Then the margin can be expressed through (rescaled) $\mathbf{w}$ and b as:

$$m = 2r = \frac{2}{\|\mathbf{w}\|}$$

Review

- Then we can formulate the *quadratic optimization problem:*

Find $\mathbf{w}$ and b such that

$$\rho = \frac{2}{\|\mathbf{w}\|} \text{ is maximized}$$

and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

Which can be reformulated as:

Find $\mathbf{w}$ and b such that

$\mathbf{\Phi}(\mathbf{w}) = \|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w}$ is minimized

and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

Review

Lagrangian of Original Problem

$$\text{Minimize } \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to } 1-y_i(\mathbf{w}^T\mathbf{x}_i+b) \leq 0 \qquad \text{for } i = 1,\ldots,n$$

The Lagrangian is

Lagrangian multipliers

$$\mathcal{L} = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^{n} \alpha_i \left(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)\right)$$

Note that $||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w}$

Setting the gradient of $\mathcal{L}$ w.r.t. **w** and b to zero, we have

$$\mathbf{w} + \sum_{i=1}^{n} \alpha_i(-y_i)\mathbf{x}_i = \mathbf{0} \quad \Rightarrow \quad \boxed{\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i}$$

$$\boxed{\sum_{i=1}^{n} \alpha_i y_i = 0}$$

$\alpha_i \geq 0$

Review

The Dual Optimization Problem

We can transform the problem to its dual

Dot product of X

$$\max. \; W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

α's → New variables
(Lagrangian multipliers)

This is a convex quadratic programming (QP) problem

– Global maximum of α_i can always be found

→well established tools for solving this optimization problem (e.g. cplex)

Note:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

General Idea: Lagrange Optimization

$$\min_w f(w)$$
$$\text{s.t. } g_i(w) \leq 0, \quad i = 1, \ldots, k$$
$$h_i(w) = 0, \quad i = 1, \ldots, l$$

1) Formulate Lagrangian function (primal problem)
$L_p(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$
2) Minimize Lagrangian wrt primal variable w: $\frac{\vartheta L_p(w, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\vartheta w} = 0$
3) Substitute the primal variable w and express Lagrangian wrt dual variables α_i, β_i*:* $L_d(\boldsymbol{\alpha}, \boldsymbol{\beta})$
4) Maximize the Lagrangian with respect to dual variables and solve for dual variables (dual problem)
ϑw
5) Recover the solution (for the primal variables) from the dual variables

Review

Note that data only appears as dot products

$$\max \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \alpha_j y_i y_j x_i x_j$$
$$\text{s.t.} \quad C \geq \alpha_i \geq 0, \sum_{i=1}^{N} \alpha_i y_i = 0$$

Since data is only represented as dot products, we need not do the mapping explicitly.

Introduce a Kernel Function (*) *K* such that:

$$K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$$

(*)Kernel function – a function that can be applied to pairs of input data to evaluate dot products in some corresponding feature space

Review

Consider the following transformation

$$\phi(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}) = (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2)$$
$$\phi(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}) = (1, \sqrt{2}y_1, \sqrt{2}y_2, y_1^2, y_2^2, \sqrt{2}y_1y_2)$$

Define the kernel function $K(\mathbf{x},\mathbf{y})$ as

$$\begin{aligned} \langle \phi(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}), \phi(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}) \rangle &= (1 + x_1y_1 + x_2y_2)^2 \\ &= K(\mathbf{x}, \mathbf{y}) \end{aligned}$$

$$K(\mathbf{x}, \mathbf{y}) = (1 + x_1y_1 + x_2y_2)^2$$

The inner product $\phi(.)\phi(.)$ can be computed by K without going through the map $\phi(.)$ explicitly!!!

Review

Kernel SVM

Change all inner products to kernel functions

For training,

Original

$$\text{max. } W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

With kernel function

$$\text{max. } W(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{subject to } C \geq \alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0$$

Review

- Linear: $K(\mathbf{x}_i,\mathbf{x}_j)= \mathbf{x}_i^T\mathbf{x}_j$
  - Mapping Φ: $\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ is $\mathbf{x}$ itself

- Polynomial of power p: $K(\mathbf{x}_i,\mathbf{x}_j)= (1+ \mathbf{x}_i^T\mathbf{x}_j)^p$
  - Mapping Φ: $\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ has $\binom{d+p}{p}$ dimensions

- Gaussian (radial-basis function): $K(\mathbf{x}_i,\mathbf{x}_j) = e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}}$
  - Mapping Φ: $\mathbf{x} \rightarrow \boldsymbol{\varphi}(\mathbf{x})$, where $\boldsymbol{\varphi}(\mathbf{x})$ is *infinite-dimensional*: every point is mapped to *a function* (a Gaussian); combination of functions for support vectors is the separator.

- Higher-dimensional space still has *intrinsic* dimensionality *d* (the mapping is not *onto*), but linear separators in it correspond to *non-linear* separators in original space.

Review

- Q1: The evolution of neuron from biology to mathematics?
- Q2: The key concepts of Perceptron and its limitations.
- Q3: Who is Vladimir N. Vapnik and what is his major contribution?
- Q4: Maximum Margin Principle and why support vector is important?
- Q5: How to compute the distance between a given data point to the decision boundary?
- Q6: What limitations the linear SVM suffered from?
- Q7: Why dual form of SVM is important?
- Q8: How to derive the dual form from the prime form of SVM?
- Q9: What the limitations the kernel method suffers from?
- Q10: the relation between Perception and SVM.
- Q11: are there other methods to address linear inseparable issue besides kernel?

Others

Review

- Q1: The evolution of neuron from biology to mathematics?
- Q2: The key concepts of Perceptron and its limitations.
- Q3: Who is Vladimir N. Vapnik and what is his major contribution?
- Q4: Maximum Margin Principle and why support vector is important?
- Q5: How to compute the distance between a given data point to the decision boundary?
- Q6: What limitations the linear SVM suffered from?
- Q7: Why dual form of SVM is important?
- Q8: How to derive the dual form from the prime form of SVM?
- Q9: What the limitations the kernel method suffers from?
- Q10: the relation between Perception and SVM.
- Q11: are there other methods to address linear inseparable issue besides kernel?

Others

Review

Q9: What the limitations the kernel method suffers from?

- No golden criterion to choose of kernel
- Kernel is specified by human, but learned from data
- Increase the data dimension, leading to high computational cost

Review

Q9: What the limitations the kernel method suffers from?

- No golden criterion to choose of kernel
- Kernel is specified by human, but learned from data
- Increase the data dimension, leading to high computational cost

Q11: are there other methods to address linear inseparable issue besides kernel?

- Dimension reduction

提纲

一．Review

二．Dimension Reduction

三．Principle Component Analysis

二、Dimension Reduction

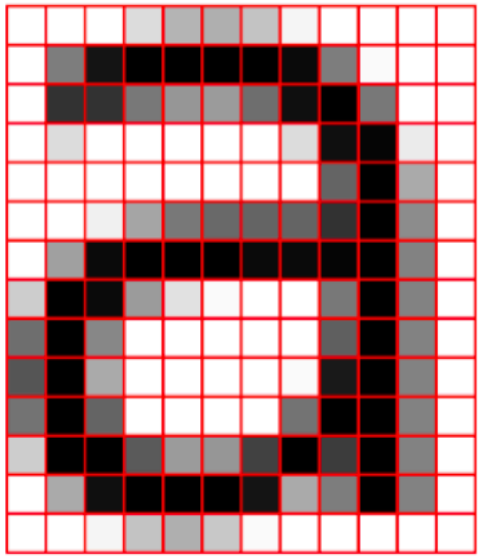
数据的高维导致所谓的维灾 ，本质上是说数据的关键特性主要分布在少量维度（属性）上，其分布于所有维度张成空间的概率接近于0。流行的Euclidean距离平等地对待每一个属性，而不加区分，同时是一种pairwise的距离，故不能真实的描述出数据之间的关系和分布特性。

二、Dimension Reduction

数据的高维导致所谓的维灾 ，本质上是说数据的关键特性主要分布在少量维度（属性）上，其分布于所有维度张成空间的概率接近于0。流行的Euclidean距离平等地对待每一个属性，而不加区分，同时是一种pairwise的距离，故不能真实的描述出数据之间的关系和分布特性。

14

14



1.0	1.0	1.0	0.9	0.6	0.6	0.6	1.0	1.0	1.0	1.0	1.0	
1.0	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	1.0	1.0	
1.0	0.2	0.2	0.5	0.6	0.6	0.5	0.0	0.0	0.5	1.0	1.0	
1.0	0.9	1.0	1.0	1.0	1.0	1.0	0.9	0.0	0.0	0.9	1.0	
1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0	
1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.4	0.0	0.5	1.0	
1.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	1.0	
0.9	0.0	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0	
0.5	0.0	0.6	1.0	1.0	1.0	1.0	1.0	0.5	0.0	0.5	1.0	
0.5	0.0	0.7	1.0	1.0	1.0	1.0	1.0	0.0	0.0	0.5	1.0	
0.6	0.0	0.6	1.0	1.0	1.0	1.0	0.5	0.0	0.0	0.5	1.0	
0.9	0.1	0.0	0.6	0.7	0.7	0.5	0.0	0.5	0.0	0.5	1.0	
1.0	0.7	0.1	0.0	0.0	0.0	0.1	0.9	0.8	0.0	0.5	1.0	
1.0	1.0	1.0	0.8	0.8	0.9	1.0	1.0	1.0	1.0	1.0	1.0	

Dimension: 数据属性的个数。上例中维度是多少？

二、Dimension Reduction

定义：**dimensionality reduction** or **dimension reduction** is the process of reducing the number of random variables/dimension under consideration by obtaining a set of principal variables – redundancy removal

High-Dimensions = Lot of Features

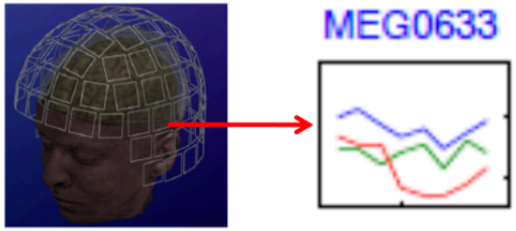
Document classification

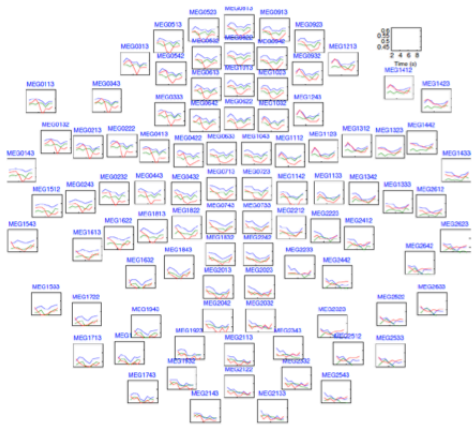
Features per document =
thousands of words/unigrams
millions of bigrams, contextual information

Or any high-dimensional image data

MEG Brain Imaging

120 locations x 500 time points
x 20 objects





二、Dimension Reduction

Good for:

- Discovering low-dimensional structural representations
- More efficient use of resource (e.g., time, memory, communication)
- Statistically, fewer dimensions -> better generalization
- Noise removal (improving data quality)
- Visualization

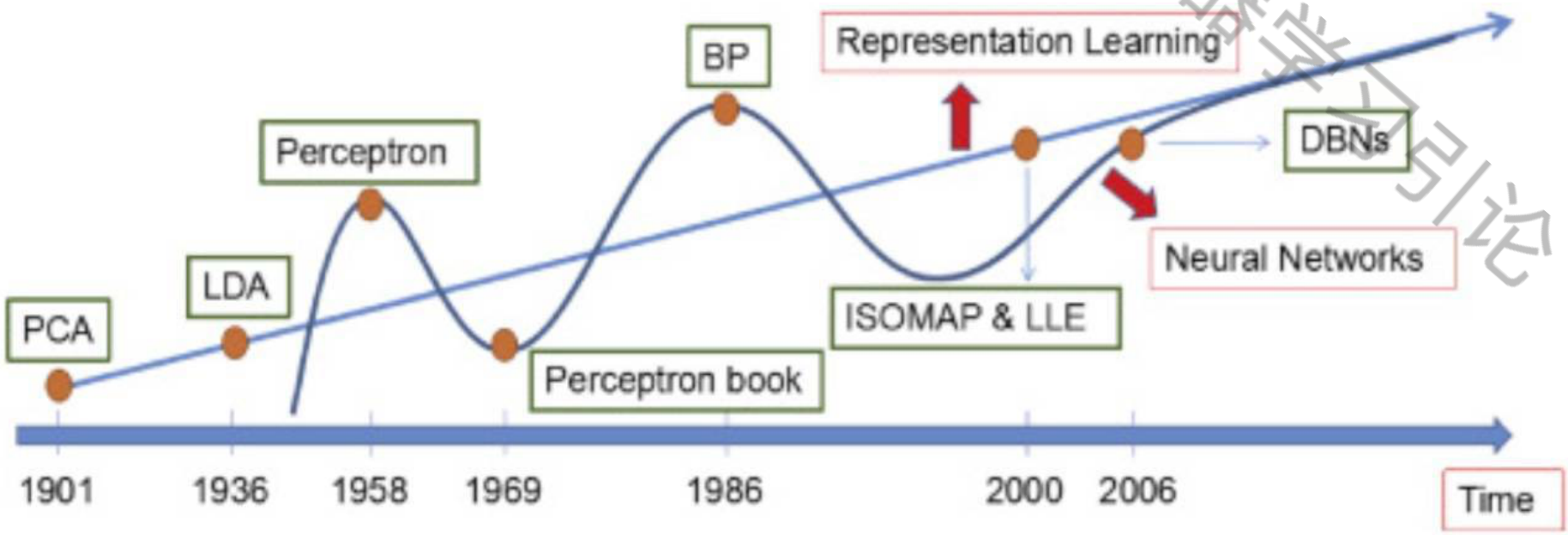
提纲

一．Review

二．Dimension Reduction

三．Principal Component Analysis

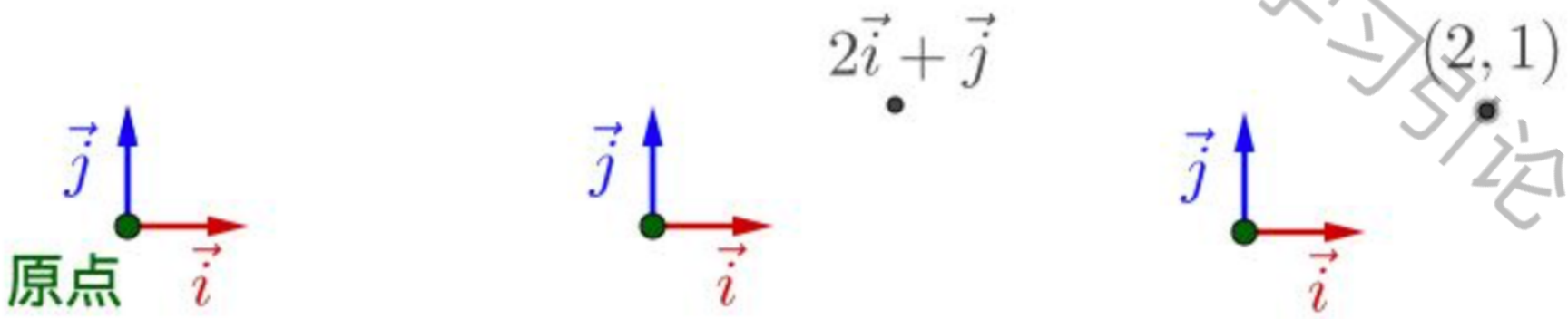
三、Principal Component Analysis



三、Principal Component Analysis

线性代数：研究**线性空间（vector space）**和其上的**线性变换**的学科。

- 线性空间：



整个二维平面上的点，都可以通过 $a\vec{i}+b\vec{j}$ 的方式来表示。

The space spanned by $\vec{i}$ and $\vec{j}$ is the two-diemsnioanl vector space.

三、Principal Component Analysis

线性空间（vector space）：

- A vector space over a field F is a set V together with **two operations** that satisfy the **eight axioms** listed below.
- The first operation, called **vector addition** or simply addition: $\mathbf{V} + \mathbf{V} \rightarrow \mathbf{V}$, takes any two vectors v and w and assigns to them a third vector which is commonly written as v + w, and called the sum of these two vectors. (Note that the resultant vector is also an element of the set V).
- The second operation, called **scalar multiplication**: $\mathbf{F} \times \mathbf{V} \rightarrow \mathbf{V}$ ， takes any scalar a and any vector v and gives another vector av. (Similarly, the vector av is an element of the set V).
- Consists of null space (0).

三、Principal Component Analysis

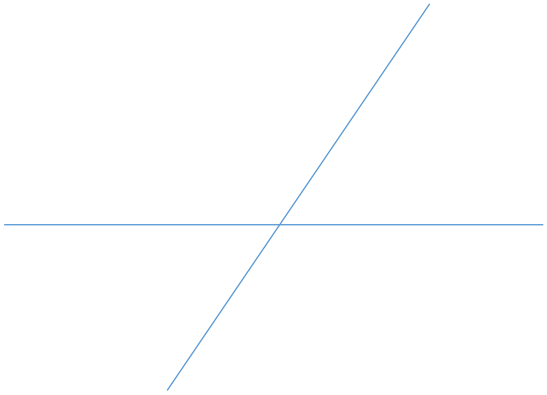
Dimensionality reduction (DR) or dimension reduction is the process of reducing the number of random variables/dimension under consideration by obtaining a set of principal variables – redundancy removal.

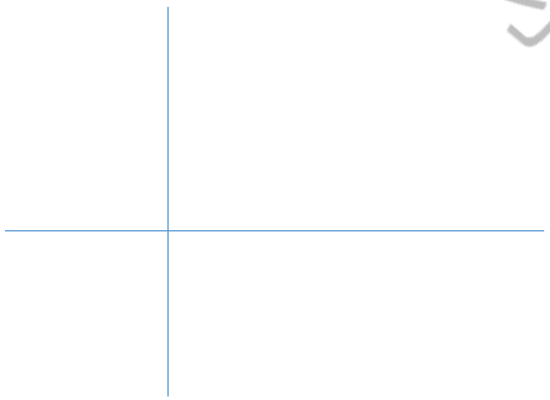
Basis：In mathematics, a set of elements (vectors) in a vector space V is called a basis, or a set of basis vectors, if the vectors are linearly independent and every vector in the vector space is a linear combination of this set.

$$\mathbf{B} = \begin{bmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \vdots \\ \mathbf{b_m} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}$$

三、Principal Component Analysis

linearly independent or orthogonal?



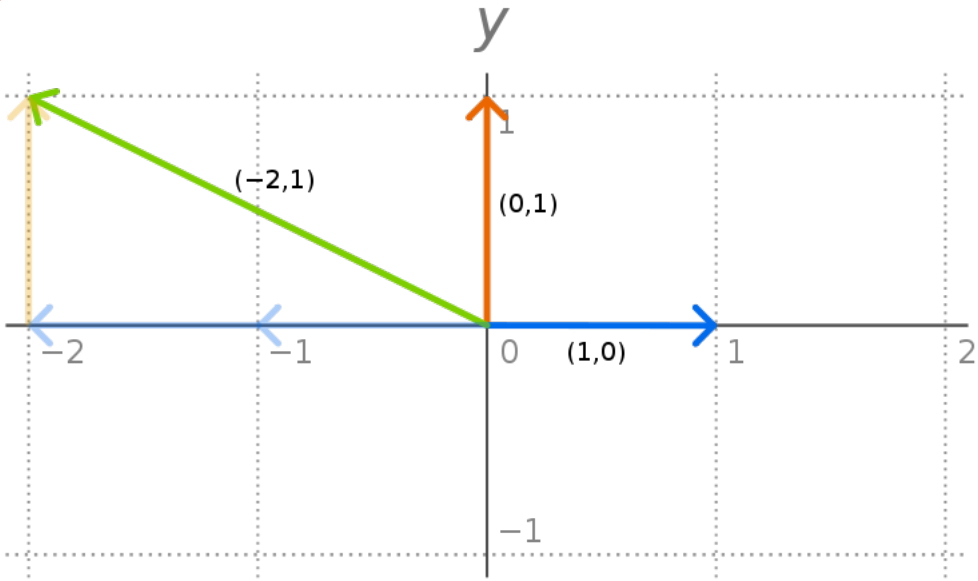


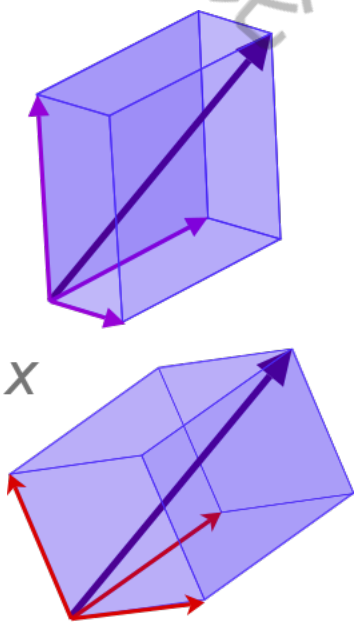
三、Principal Component Analysis

Dimensionality reduction (DR) or dimension reduction is the process of reducing the number of random variables/dimension under consideration by obtaining a set of principal variables – redundancy removal.

Basis：In mathematics, a set of elements (vectors) in a vector space V is called a basis, or a set of basis vectors, if the vectors are linearly independent and every vector in the vector space is a linear combination of this set.

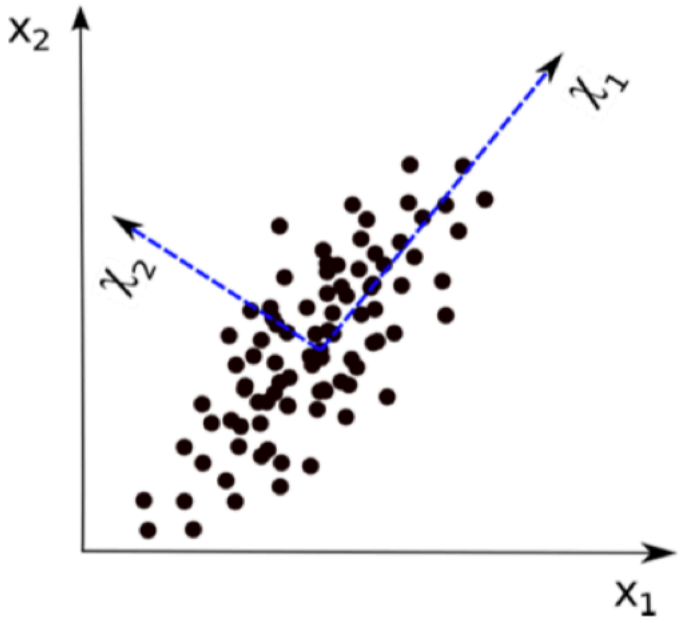
DR could be achieved by seeking the basis of a give data set!

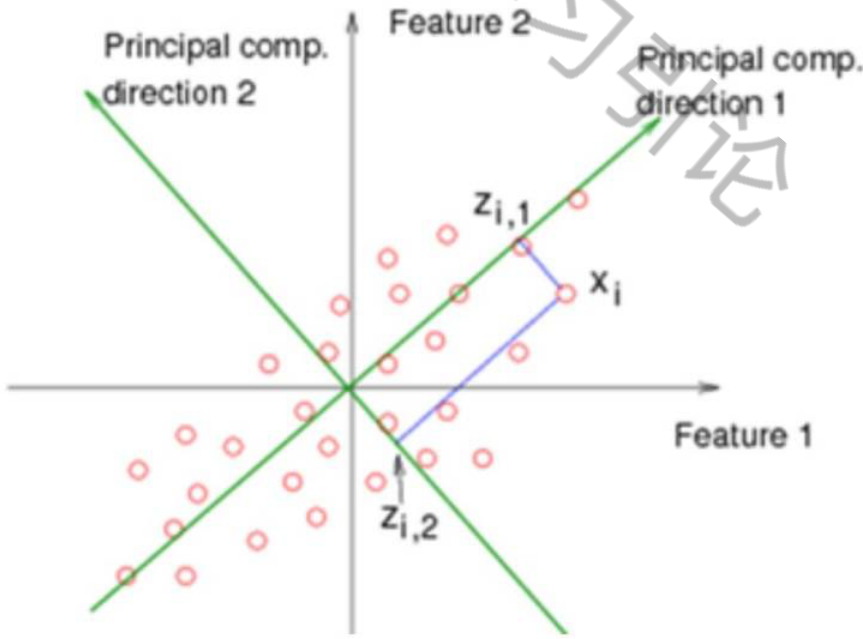




三、Principal Component Analysis

Component = orthogonal basis





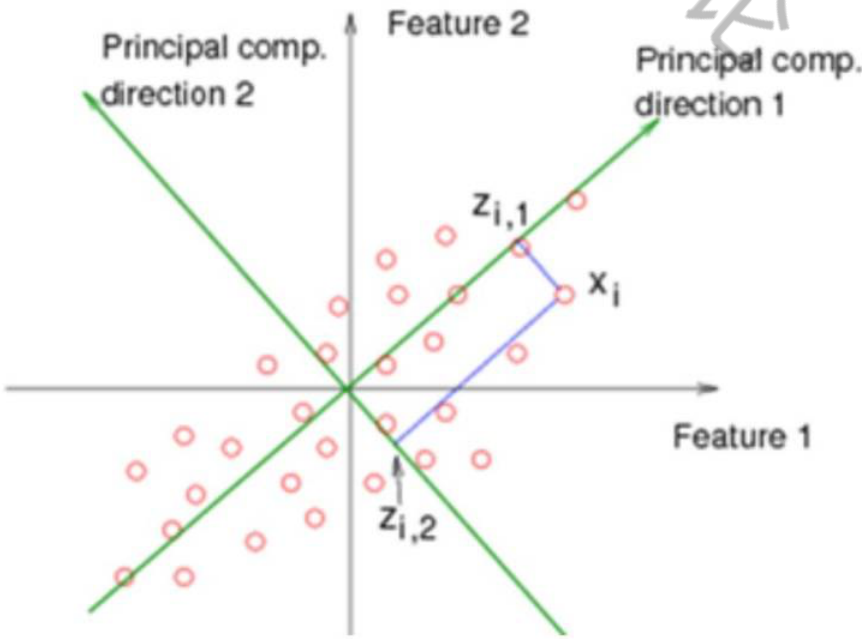
三、Principal Component Analysis

PCA aims to seek the components W from a given data set X, i.e.,

$$Y=W^TX,$$

where W is the set of components (projection matrix), and Y is the reduced representation of X.

How?



三、Principal Component Analysis

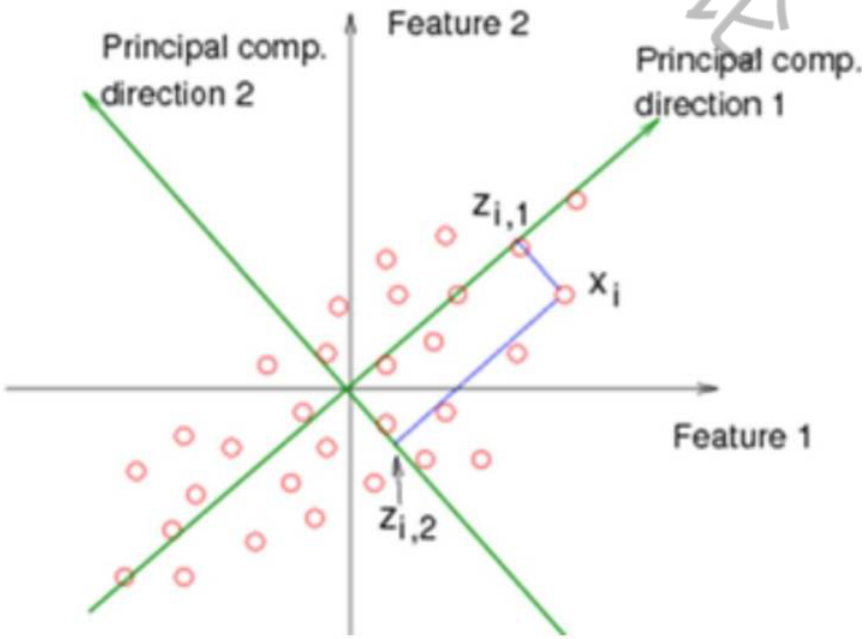
PCA aims to seek the components W from a given data set X, i.e.,

$$Y=W^TX,$$

where W is the set of components (projection matrix), and Y is the reduced representation of X.

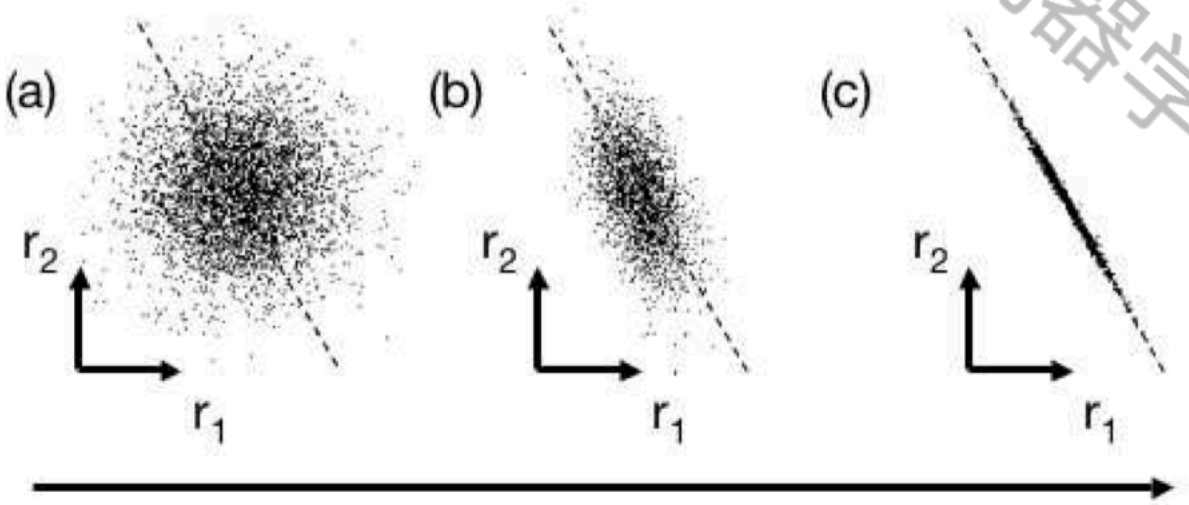
How?

The intrinsic assumption of DR is that the data contains redundancy!



三、Principal Component Analysis

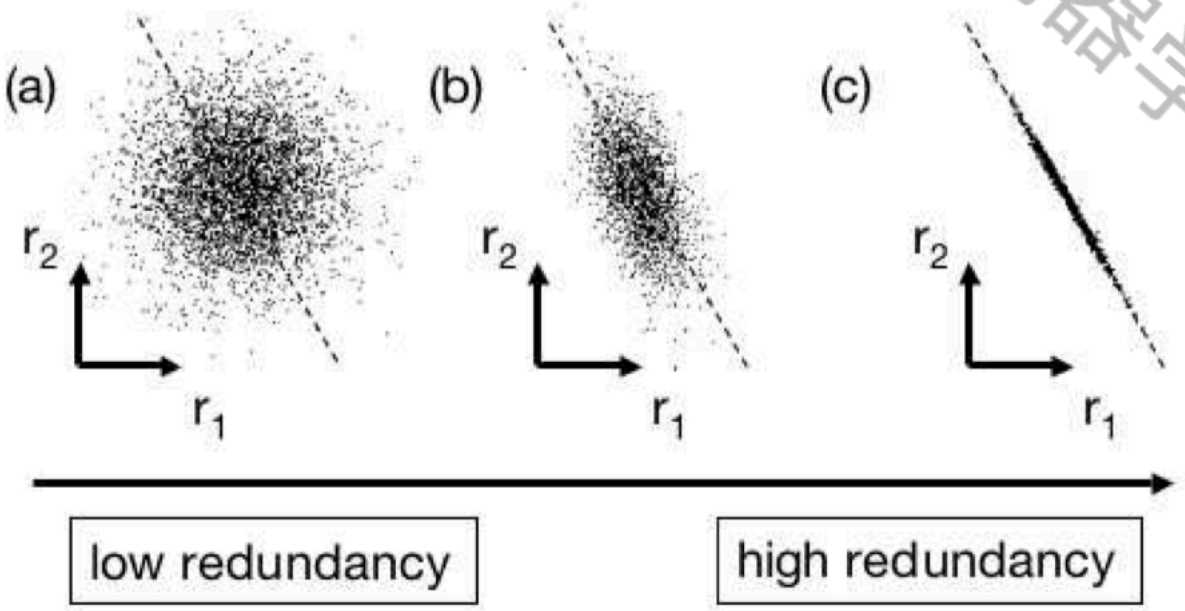
What is redundancy?



Higher or lower?

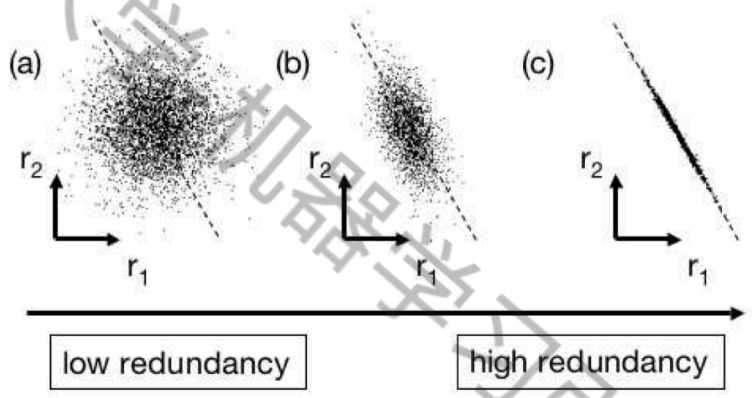
三、Principal Component Analysis

What is redundancy?



三、Principal Component Analysis

How to measure redundancy in mathematics?



- Covariance:

If the entries in the column vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

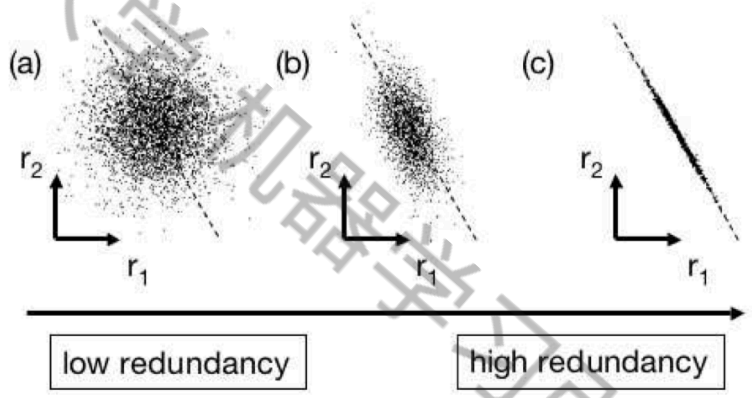
are random variables, each with finite variance, then the covariance matrix Σ is the matrix whose (*i*, *j*) entry is the covariance

$$\Sigma_{ij} = \operatorname{cov}(X_i, X_j) = \operatorname{E}[(X_i - \mu_i)(X_j - \mu_j)] = \operatorname{E}[X_i X_j] - \mu_i \mu_j$$

where the operator E denotes the expected (mean) value of its argument, and

$$\mu_i = \operatorname{E}(X_i)$$

三、Principal Component Analysis



What is redundancy in mathematics?

- Covariance:

If the entries in the column vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

are random variables, each with finite variance, then the covariance matrix Σ is the matrix whose (*i*, *j*) entry is the covariance

$$\Sigma_{ij} = \operatorname{cov}(X_i, X_j) = \operatorname{E}[(X_i - \mu_i)(X_j - \mu_j)] = \operatorname{E}[X_i X_j] - \mu_i \mu_j$$

where the operator E denotes the expected (mean) value of its argument, and

$$\mu_i = \operatorname{E}(X_i)$$

- $\sum_{ij} = 0$ if and only if i and j are entirely uncorrelated.
- Otherwise, i and j are correlated.

Correlated=redundant!

Futher reading: In fact, the variances $\sum_{ij}$ also defines the signal-to-noise ratio.

三、Principal Component Analysis

Let the data set X be with zero mean, then define

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^{\mathbf{T}}\mathbf{X}$$

i.e.,

$$\mathrm{C} = \frac{1}{n-1}\begin{bmatrix} \boldsymbol{x}^2{}_1 & \cdots & \boldsymbol{x}_1\boldsymbol{x}_n \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_n\boldsymbol{x}_1 & \cdots & \boldsymbol{x}^2{}_n \end{bmatrix}$$

Some properties of $\mathbf{C_X}$:

- $\mathbf{C_X}$ is a square symmetric
- The diagonal terms of $\mathbf{C_X}$ are the *variance* of particular measurement types.
- The off-diagonal terms of $\mathbf{C_X}$ are the *covariance* between measurement types.

$\mathbf{C_X}$ captures the correlations between all possible pairs of measurements. The correlation values reflect the noise and redundancy in our measurements.

- In the diagonal terms, by assumption, large (small) values correspond to interesting dynamics (or noise).
- In the off-diagonal terms large (small) values correspond to high (low) redundancy.

三、Principal Component Analysis

As the covariance defines the redundancy, then one could remove the redundancy in low dimensional space by diagonalizing the covariance matrix.

投影：$\mathbf{w}^T \boldsymbol{x}$

方差：$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T \boldsymbol{x}_i)^2 = \mathbf{w}^T S \mathbf{w}$

$$S = \frac{1}{n}\sum_i \boldsymbol{x}_i \boldsymbol{x}_i^T$$

拉格朗日乘数法：

$$L = \mathbf{w}^T S \mathbf{w} + \lambda(1 - \mathbf{w}^T \mathbf{w})$$

$$\frac{\partial L}{\partial \mathbf{w}} = 2S\mathbf{w} - 2\lambda\mathbf{w}$$

$$S\mathbf{w} = \lambda\mathbf{w}$$

最大方差：

$$\max_{\mathbf{w}} \quad \mathbf{w}^T S \mathbf{w}$$

$$s.t. \quad \|\mathbf{w}\| = 1$$

方差：

$$\mathbf{w}^T S \mathbf{w} = \mathbf{w}^T \lambda \mathbf{w} = \lambda$$

三、Principal Component Analysis

As the covariance defines the redundancy, then one could remove the redundancy in low dimensional space by diagonalizing the covariance matrix.

投影：$\mathbf{w}^T\boldsymbol{x}$

方差：$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{w}^T\boldsymbol{x}_i)^2 = \mathbf{w}^T S\mathbf{w}$

$$S = \frac{1}{n}\sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T$$

拉格朗日乘数法：

$$L = \mathbf{w}^T S\mathbf{w} + \lambda(1 - \mathbf{w}^T\mathbf{w})$$

$$\frac{\partial L}{\partial \mathbf{w}} = 2S\mathbf{w} - 2\lambda\mathbf{w}$$

$$S\mathbf{w} = \lambda\mathbf{w}$$

最大方差：

$$\max_{\mathbf{w}} \quad \mathbf{w}^T S\mathbf{w}$$

$$s.t. \quad \|\mathbf{w}\| = 1$$

Why

方差：

$$\mathbf{w}^T S\mathbf{w} = \mathbf{w}^T\lambda\mathbf{w} = \lambda$$

三、Principal Component Analysis

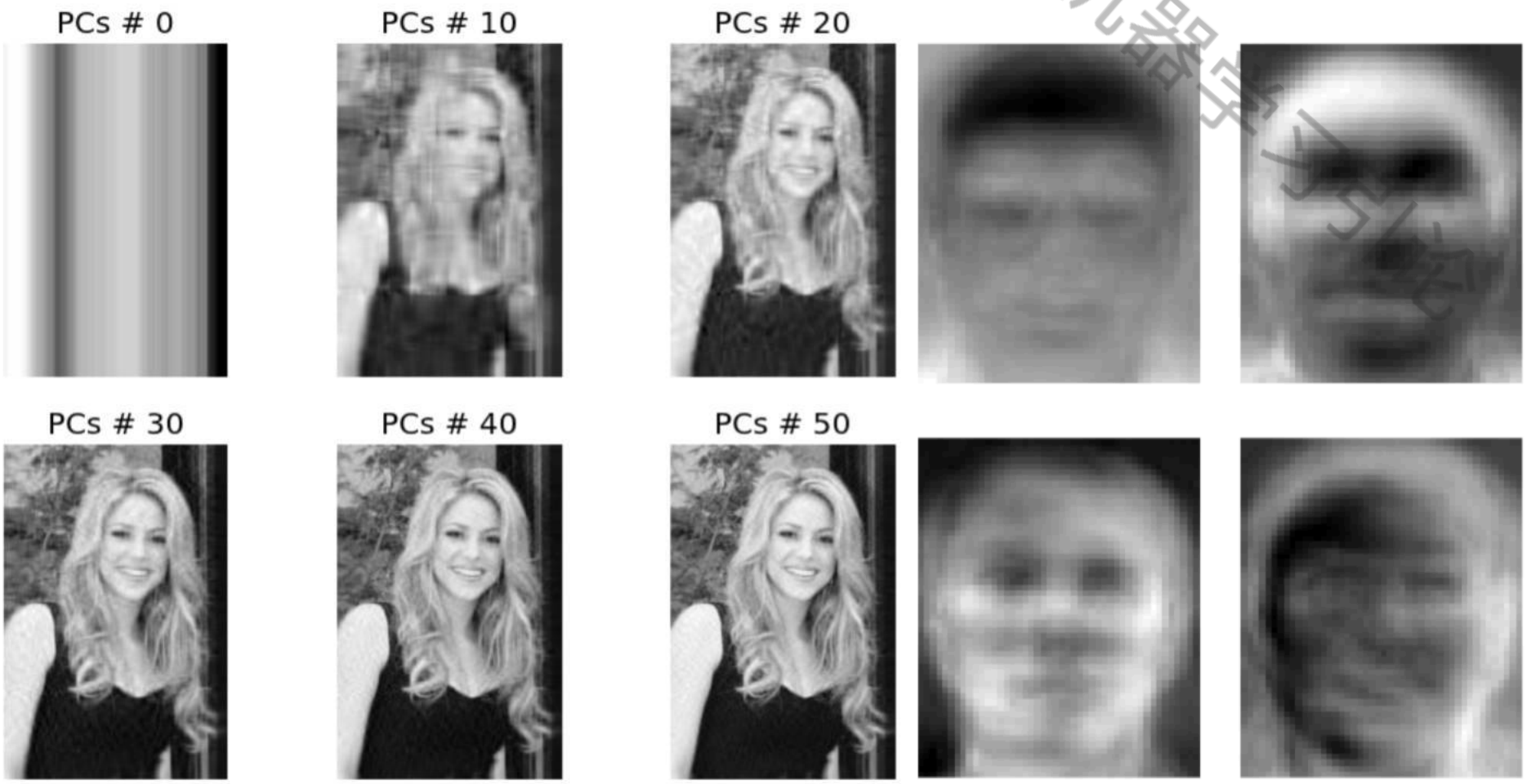
- Select a normalized direction in m-dimensional space along which the variance in X is maximized. Save this vector as $\mathbf{w}_1$.
- Find another direction along which variance is maximized, however, because of the orthonormality condition, restrict the search to all directions perpendicular to all previous selected directions. Save this vector as $\mathbf{w}_i$
- Repeat this procedure until m vectors are selected.

Simply:

- Computing and using *m* largest eigenvectors of X^TX as the column of W.

Q: beside performing ED on X^TX, is there other methods to obtain the principal components?

三、Principal Component Analysis

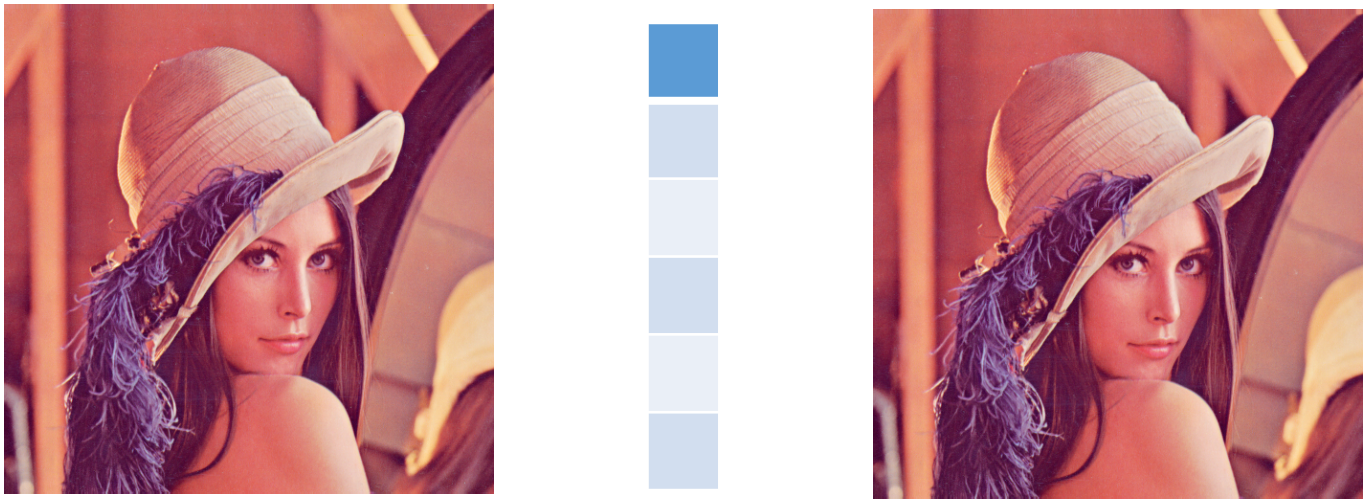


三、Principal Component Analysis

View 1: Redundancy removal

- Covariance measures redundancy
- thus DR could be achieved by diagonalizing the covariance matrix
- leading to the ED on X^TX

View 2: minimizing reconstruction error/description length.



Input low dimensional rep. Reconstruction

三、Principal Component Analysis

View 2: minimizing reconstruction error/description length.

正交基：$\mathbf{u}_1, \cdots, \mathbf{u}_D$

原始数据：$\boldsymbol{x}_i = \sum_{j=1}^{D} \alpha_{ij} \mathbf{u}_j$

基坐标：$\alpha_{ij} = \mathbf{u}_j^T \boldsymbol{x}_i$

降维重建：$\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d} \alpha_{ij} \mathbf{u}_j$

减一维重建：

$d = D - 1$

对应最小特征值：

$\mathbf{w}_D^T S \mathbf{w}_D = \lambda_D$

三、Principal Component Analysis

View 2: minimizing reconstruction error/description length.

正交基： $\mathbf{u}_1, \cdots, \mathbf{u}_D$

原始数据： $\boldsymbol{x}_i = \sum_{j=1}^{D} \alpha_{ij} \mathbf{u}_j$

基坐标： $\alpha_{ij} = \mathbf{u}_j^T \boldsymbol{x}_i$

降维重建： $\hat{\boldsymbol{x}}_i = \sum_{j=1}^{d} \alpha_{ij} \mathbf{u}_j$

$$\frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 = \frac{1}{n}\sum_{i=1}^{n} \|\sum_{j=1}^{D} \alpha_{ij}\mathbf{u}_j - \sum_{j=1}^{d} \alpha_{ij}\mathbf{u}_j\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} \|\sum_{j=d+1}^{D} \alpha_{ij}\mathbf{u}_j\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=d+1}^{D} \alpha_{ij}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=d+1}^{D} \mathbf{u}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \mathbf{u}_j$$

$$= \sum_{j=d+1}^{D} \mathbf{u}_j^T S \mathbf{u}_j$$ 等价方差最小

Taking Home:

- What means Principal?
- What means Components? And Principal Components.
- The definition of basis and its relation with component.
- Redundancy and how to measure it.
- Covariance matrix.
- Why PCA diagonalizes the covariance matrix, what are denoted by off/on-diagonal entries of the covariance matrix?
- How to reconstruct a data point for a given PCs?

Test Questions

Q1: Why the variances $\sum_{ij}$ also defines the signal-to-noise ratio? And the properties of SNR w.r.t. $\sum_{ij}$.
Q2: Beside performing ED on X^TX, is there other way to obtain the principal components?
Q3: Can PCA handle the data drawn from multiple subspace?
Q4: PCA is a unsupervised dimension reduction method, which may suffer from what problem or limitations?
Q5: What distance is adopted by PCA to measure the relation among data points? Could such a measurement solve the linear inseparable issue? If Yes/NO, why?

And so on…

Next

Q1: Why the variances $\sum_{ij}$ also defines the signal-to-noise ratio? And the properties of SNR w.r.t. $\sum_{ij}$.

Q2: Beside performing ED on X^TX, is there other way to obtain the principal components?

Q3: Can PCA handle the data drawn from multiple subspace? Why?

Q4: PCA is a unsupervised dimension reduction method, which may suffer from what problem or limitations?

Q5: What distance is adopted by PCA to measure the relation among data points? Could such a measurement solve the linear inseparable issue? If Yes/NO, why?

Multiple subspace dimension reduction

+

Supervised dimension reduction

Q&A

THANKS!