Midterm Review

Mathematical Foundations for the Information Age

Peking University

October 24th, 2024

Contents

Contents

Dimensionality Reduction

Dimensionality reduction is a very important technique in many areas.

- Data Compression
- Recommendation System
- Computer Vision
- ...

Review

SVD gives best rank-k approximations.

Theorem

For any matrix B of rank at most k, we have

- $||A - A_k||_F \leq ||A - B||_F$
- $||A - A_k||_2 \leq ||A - B||_2$

Lemma

The rows of A_k are the projections of the rows of A onto the subspace V_k spanned by the first k singular vectors of A.

Example

Given 2 points $(2, 0), (0, 1)$, find the best 1-dimensional subspace V_1?

Example

Given 2 points $(2,0),(0,1)$, find the best 1-dimensional subspace V_1?

Run SVD on $A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ and let $V_1 = \text{span}(\{\boldsymbol{v}_1\}) = \{\lambda \boldsymbol{v}_1 | \lambda \in \mathbb{R}\}$

Example

Given 2 points $(2,0),(0,1)$, find the best 1-dimensional subspace V_1?

Run SVD on $A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ and let $V_1 = \text{span}\left(\{\boldsymbol{v}_1\}\right) = \{\lambda \boldsymbol{v}_1 | \lambda \in \mathbb{R}\}$

$\boldsymbol{v}_1 = \{(1,0)\}, V_1 = \text{span}\left(\{\boldsymbol{v}_1\}\right) = \{(x_1, 0) | x_1 \in \mathbb{R}\}$

Example

Given 2 points $(2,0),(0,1)$, find the best 1-dimensional subspace V_1?

Run SVD on $A=\begin{bmatrix}2 & 0\\0 & 1\end{bmatrix}$ and let $V_1=\text{span}\left(\{\boldsymbol{v}_1\}\right)=\{\lambda\boldsymbol{v}_1|\lambda\in\mathbb{R}\}$

$\boldsymbol{v}_1=\{(1,0)\}, V_1=\text{span}\left(\{\boldsymbol{v}_1\}\right)=\{(x_1,0)|x_1\in\mathbb{R}\}$

But obviously $(2,0)$ and $(0,1)$ is on the line $x_1+2x_2=2$.

Why are the results different?

Example

Given 2 points $(2,0),(0,1)$, find the best 1-dimensional subspace V_1?

Run SVD on $A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$ and let $V_1 = \text{span}\left(\{\boldsymbol{v}_1\}\right) = \{\lambda \boldsymbol{v}_1 | \lambda \in \mathbb{R}\}$

$\boldsymbol{v}_1 = \{(1,0)\}, V_1 = \text{span}\left(\{\boldsymbol{v}_1\}\right) = \{(x_1, 0) | x_1 \in \mathbb{R}\}$

But obviously $(2,0)$ and $(0,1)$ is on the line $x_1 + 2x_2 = 2$.
Why are the results different?

Remark

A subspace must contain the zero vector $\mathbf{0}$.

Main Question

But sometimes the coordinate system is chosen arbitrarily. Requiring the space after dimension reduction to include the origin is not necessary.

Main Question

But sometimes the coordinate system is chosen arbitrarily. Requiring the space after dimension reduction to include the origin is not necessary. So how can we find a k-dimensional space(may not include the origin) that best fits the data points?

Remark

Here 'best fitting the data points' means 'minimizing the sum of the squared perpendicular distances to these points'.

Centering Data

Intuition

Find a point in the best k-dimensional space. Let it be the new origin point. Then use SVD in new coordinate system.

Centering Data

Intuition

Find a point in the best k-dimensional space. Let it be the new origin point. Then use SVD in new coordinate system.

How to find a point that will always in the best k-dimensional space?

Centering Data

Intuition

Find a point in the best k-dimensional space. Let it be the new origin point. Then use SVD in new coordinate system.

How to find a point that will always in the best k-dimensional space?

Example: Linear Regression

Find the best $y = kx + b$ best fitting $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.
Solution: $\hat{b} = \bar{y} - \hat{k}\bar{x}$. $(\bar{x}, \bar{y})$ always on the line.

Centering Data

Intuition

Find a point in the best k-dimensional space. Let it be the new origin point. Then use SVD in new coordinate system.

How to find a point that will always in the best k-dimensional space?

Example: Linear Regression

Find the best $y = kx + b$ best fitting $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$.
Solution: $\hat{b} = \bar{y} - \hat{k}\bar{x}$. $(\bar{x}, \bar{y})$ always on the line.

Centering data:

- Subtracting the centroid(the coordinate-wise average) of the data from each data point.

Main Theorem

Theorem

The best fitting k-dimension space must pass through the centroid of the points.

Affine Space

How to represent a general k-dimension space in $\mathbb{R}^d$?

Affine Space

How to represent a general k-dimension space in $\mathbb{R}^d$?

Definition

A k-dimensional affine space in $\mathbb{R}^d$ is a set of the form

$$\{\boldsymbol{v}_0 + \sum_{i=1}^{k} c_i \boldsymbol{v}_i | c_1, \ldots, c_k \in \mathbb{R}\}$$

where $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ are pairwise otrhonormal and $\boldsymbol{v}_0 \perp \boldsymbol{v}_i = 0$

$\text{Span}\left(\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}\right)$ forms a k-dimensional subspace, and $\boldsymbol{v}_0$ acts as an offset.

Affine Space

How to represent a general k-dimension space in $\mathbb{R}^d$?

Definition

A k-dimensional affine space in $\mathbb{R}^d$ is a set of the form

$$\{\boldsymbol{v}_0 + \sum_{i=1}^{k} c_i \boldsymbol{v}_i | c_1, \ldots, c_k \in \mathbb{R}\}$$

where $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k)$ are pairwise otrhonormal and $\boldsymbol{v}_0 \perp \boldsymbol{v}_i = 0$

Span $(\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\})$ forms a k-dimensional subspace, and $\boldsymbol{v}_0$ acts as an offset.

Example

A line $y = kx + b$ in $\mathbb{R}^2$ can be represented as

$$(x, y) \in \{(\frac{-bk}{k^2+1}, \frac{b}{k^2+1}) + \lambda(1, k) | \lambda \in \mathbb{R}\}$$

Affine Space

The distance of a point $\boldsymbol{a}_i \in \mathbb{R}^d$ to the k-dimensional affine space $W = \{\boldsymbol{v}_0 + \sum_{i=1}^k c_i \boldsymbol{v}_i | c_1, \dots, c_k \in \mathbb{R}\}$ satisfies

$$\begin{aligned} dist(\boldsymbol{a}_i, W)^2 &= dist(\boldsymbol{a}_i - \boldsymbol{v}_0, \text{span}(\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_k\}))^2 \\ &= |\boldsymbol{a}_i - \boldsymbol{v}_0|^2 - \sum_{j=1}^k ((\boldsymbol{a}_i - \boldsymbol{v}_0) \cdot \boldsymbol{v}_j)^2 \\ &= |\boldsymbol{a}_i - \boldsymbol{v}_0|^2 - \sum_{j=1}^k (\boldsymbol{a}_i \cdot \boldsymbol{v}_j)^2 \end{aligned}$$

Theorem

Theorem

The best fitting k-dimension space must pass through the centroid of the points.

Proof

Proof

Consider a new coordinate system with every point subtracting $\bar{\boldsymbol{a}} = \sum_{i=1}^{n} \boldsymbol{a}_i / n$. In this new coordinate system, data point $\boldsymbol{a}_i$ moves to $\boldsymbol{a}_i' = \boldsymbol{a}_i - \bar{\boldsymbol{a}}$

For a k-dimensional space $W = \{\boldsymbol{v}_0 + \sum_{i=1}^{k} c_i \boldsymbol{v}_i | c_1, \ldots, c_k \in \mathbb{R}\}$,

$$\begin{aligned}\sum_{i=1}^{n} dist(\boldsymbol{a}_i', W)^2 &= \sum_{i=1}^{n} (|\boldsymbol{a}_i' - \boldsymbol{v}_0|^2 - \sum_{j=1}^{k} (\boldsymbol{a}_i' \cdot \boldsymbol{v}_j)^2) \\ &= n\boldsymbol{v}_0^2 - 2(\sum_{i=1}^{n} \boldsymbol{a}_i') \cdot \boldsymbol{v}_0 + \sum_{i=1}^{n} |\boldsymbol{a}_i'|^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} (\boldsymbol{a}_i' \cdot \boldsymbol{v}_j)^2\end{aligned}$$

Maximized when $\boldsymbol{v}_0 = \sum_{i=1}^{n} \boldsymbol{a}_i'/n = \boldsymbol{0}$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k$ are the first k singular vectors for $\boldsymbol{A} = [\boldsymbol{a}_1', \cdots, \boldsymbol{a}_n']^T$

Recap

How to find the best fitting k-dimension space given some data points?

- Subtracting the centroid of the data from each data point.
- Do SVD on new data to get the best rank k approximation.

Theorem

The best fitting k-dimension space must pass through the centroid of the points.

Contents

Instructions

- **Thursday, October 31st 15:10-17:10**
- **Room 503, No.3 Teaching Building**
- **Closed-book exam**
- No paper materials or electronic devices are allowed. You need to take your **student ID card** to verify your identity.
- Contents: Basic probability inequalities, High dimensional geometry and Singular value decomposition.

Instructions

- This exam consists of about 4 problems.
- All problems are given in English. You can raise your hand to ask TA to translate certain terms that you do not understand.
- You are allowed to write your answers in Chinese, English, or a combination of both languages.
- Please clearly indicate the problem numbers before your answers.
- Please manage your time wisely.

Focus Points

- The statement of the theorems and facts learned.
- Formal proofs and intuitions of the theorems and facts.
- Applications of the theorems and facts learned.
- Problems in homework.

Basic Probability Inequalities

- Markov Inequality
- Chebyshev Inequality
- Union bound

Remark

You need to be familiar with the statements, conditions and applications of these inequalities. Before applying them, remember to check the conditions.

High Dimensional Geometry

- Properties for unit ball in $\mathbb{R}^d$.
  - Volume and surface area.
  - Concentration properties.
  - Relations with high dimensional Gaussian random variables. (How to sample uniformly in the unit ball?)
- Johnson-Lindenstrauss Lemma.

Remark

For the proof of Johnson-Lindenstrauss Lemma, you can apply Gaussian Annulus Theorem when necessary. The proof of Gaussisan Annulus Theorem is not required.

More on Johnson-Lindenstrauss Lemma

- Approximate norm:

$$\|\mathbf{\Pi} \boldsymbol{x}_i\|_2 \approx \|\boldsymbol{x}_i\|_2.$$

- Approximate the square of the norm:

$$\|\mathbf{\Pi} \boldsymbol{x}_i\|_2 \leq (1+\epsilon)\|\boldsymbol{x}_i\|_2 \Rightarrow \|\mathbf{\Pi} \boldsymbol{x}_i\|_2^2 \leq (1+\epsilon)^2\|\boldsymbol{x}_i\|_2^2 \leq (1+3\epsilon)\|\boldsymbol{x}_i\|_2^2.$$

- Approximate the inner product:

$$2\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle = \|\boldsymbol{x}_i + \boldsymbol{x_j}\|_2^2 - \|\boldsymbol{x}_i\|_2^2 - \|\boldsymbol{x}_j\|_2^2.$$

  We can increase the dimension of projection subspace to keep the norm of $\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{x}_i + \boldsymbol{x}_j$ approximately at the same time.

- You can refer to Problem 3(1), 3(2) in Homework #3 for a more detailed discussion of these properties.

Singular Value Decomposition

- Definition and geometric interpretation.
- Best fit subspace and "greedy" construction.
- Low rank approximations: F-norm, 2-norm.
- Left singular vectors and its properties.
- Relations with the eigen decomposition of $\boldsymbol{A}^\top \boldsymbol{A}$.
- Power method.
- Centering data.

Singular Value Decomposition

More on best fit subspace and "greedy" construction

- Definition of best fit subspace:

$$\max_{\boldsymbol{v}_1 \perp \boldsymbol{v}_2, \|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\| = 1} \|\boldsymbol{A}\boldsymbol{v}_1\|_2^2 + \|\boldsymbol{A}\boldsymbol{v}_2\|_2^2.$$

- Under this definition, the value doesn't depend on the basis given the fixed subspace. This is necessary in the proof of optimum for "greedy" construction. (Exercise: Find out why it's necessary!)
- Exercise: If we define the best fit subspace as

$$\max_{\boldsymbol{v}_1 \perp \boldsymbol{v}_2, \|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\| = 1} \|\boldsymbol{A}\boldsymbol{v}_1\|_2 + \|\boldsymbol{A}\boldsymbol{v}_2\|_2,$$

  will the "greedy" construction work?

More on left singular vectors

- Define the first and second singular vectors for matrix $\boldsymbol{A}$ as $\boldsymbol{v}_1, \boldsymbol{v}_2$, then we have $\boldsymbol{A}\boldsymbol{v}_1 \perp \boldsymbol{A}\boldsymbol{v}_2$.
- Consider

$$f(\theta) := \|\boldsymbol{A}(\cos\theta\,\boldsymbol{v}_1 + \sin\theta\,\boldsymbol{v}_2)\|_2^2 \leq \sigma_1^2.$$

- We have

$$f(0) = \sigma_1^2 \Rightarrow f'(0) = 0.$$